DOCTEUR EN SCIENCES

Développement critique de méthodes d'analyse statistique de l'expression différentielle de gènes et de groupes de gènes, mesurées sur damiers à ADN.

Berger, Fabrice

Award date:
2009

Awarding institution:
Universite de Namur

Facultés Universitaires
Notre-Dame de la Paix

Faculté des Sciences
Département de Biologie

Développement critique de méthodes d'analyse statistique de l'expression différentielle de gènes et de groupes de gènes, mesurée sur damiers à ADN

Dissertation présentée par
Fabrice Berger
en vue de l'obtention du grade
de Docteur en Sciences

Composition du jury :
Mauro Delorenzi (SIB, Lausanne, CH)
Eric Depiereux (Promoteur, FUNDP)
Christophe Lambert (BioXPR, Namur)
Marcel Remon (FUNDP)
Jean-Louis Ruelle (GlaxoSmithKline Biologicals, Rixensart)

2009

Rempart de la Vierge, 13
B - 5000 Namur (Belgique)

Imprimé en Belgique
ISBN : 978-2-87037-651-5
Dépôt légal: D / 2009 / 1881 / 38

Facultés Universitaires Notre-Dame de la Paix
Faculté des Sciences
Rue de Bruxelles, 61, B-5000, Namur

Développement critique de méthodes d'analyse statistique de l'expression différentielle de gènes et de groupes de gènes, mesurée sur damiers à ADN

Par Fabrice Berger

Les puces à ADN permettent de mesurer le niveau d'expression de chacun des gènes du génome humain (ou d'autres organismes). Leur coût limite cependant le nombre de mesures réalisées, et les analyses statistiques traditionnelles sont peu performantes. Différentes méthodes ont été mises au point pour optimiser l'analyse de l'expression différentielle. Les plus représentatives sont décrites, pour faciliter la compréhension de la problématique, et des avantages et faiblesses présentées par les méthodes actuelles.

Notre première démarche vise à guider l'analyse sur base d'informations biologiques ou empiriques. Pour améliorer les performances de l'analyse, nous proposons de partager de l'information entre les gènes, groupés sur base d'un critère valide. La méthode *window t-test* a été mise au point en utilisant la relation empirique entre la variabilité et le niveau d'expression. Le *window t-test* dépends uniquement du nombre de mesures disponibles, qui détermine le nombre de gènes utilisés. Les performances obtenues avec un nombre limité de mesures (<=5) sont égales, voire supérieures, aux meilleurs méthodes actuelles [1].

Plusieurs critères biologiques permettent de définir des groupes de gènes (voie métabolique, localisation chromosomique...). Les méthodes actuelles d'analyse de groupes considèrent des hypothèses différentes. Notre démarche généralise l'étude des groupes par la question Q_0 : « L'expression des membres du groupe est-elle différente entre deux conditions ? ». Nous avons conçu la méthode *FAERI* pour répondre à cette question en regard de trois critères : la corrélation des gènes, le niveau d'expression individuel, et la direction de la réponse (sur- ou sous-expression). *FAERI* utilise la procédure *ANOVA-2*, précédée par deux étapes de réduction des données (standardisation Z, réduction directionnelle). La méthode *ANOVA-2* est la plus performante pour l'étude de groupes dont les membres sont activés et réprimés. La méthode *FAERI* est la plus appropriée pour tous les types de groupes testés.

Nos conclusions et notre démarche ont été automatisées, et matérialisées sous la forme du logiciel *PEGASE*. L'évaluation d'un *consensus* au départ des résultats de plusieurs méthodes d'analyse y est proposé, pour assurer à l'utilisateur l'obtention de résultats de bonne qualité, quelle que soit la méthode optimale. Plusieurs méthodes d'analyse sont proposées pour l'étude des gènes et des groupes, et les performances peuvent être évaluées sur base de données théoriques ou empiriques. *PEGASE* est utilisé par le serveur *PHOENIX* pour analyser les données [2].

1. Berger F., De Hertogh B., Pierre M., Gaigneaux A. & Depiereux E. The "Window t-test": a simple and powerful approach to detect differentially expressed genes in microarray datasets. *Cent. Eur. J. Biol.*, 2008, 3, 327-344.

2. Berger F., De Hertogh B., Bareke E., Pierre M., Gaigneaux A. & Depiereux E. PHOENIX: a web-interface for (re)analyses of microarray data. *Cent. Eur. J. Biol.*, 2009, 4(4) : 603 : 618.

Facultés Universitaires Notre-Dame de la Paix
Faculté des Sciences
Rue de Bruxelles, 61, B-5000, Namur

Development of statistical methods for differential expression analysis of genes and gene-sets from DNA microarray data

By Fabrice Berger

DNA microarrays allow to study the expression profile of the whole genome of an organism. This technology is quite expensive, and the number of tested samples is often limited (at most 5 replicates). In those conditions, statistical tests are associated to bad performances. Various methods have been developed to optimize differential expression analysis. We describe a set of methods, to provide a comprehensive view of various approaches, of their main advantages and limitations.

Our first objective is to guide the analysis of differential expression, at the gene-level, by using biological or empirical informations. To improve the performances of statistical tests, we propose to share information across genes, gathered using an appropriate *criteria*. The *window* t-test has been developed following this strategy, to use the empirical relationship between variability and mean expression level. The *window* t-test only depends on the number of replicates, that defines the number of probesets used to compute variance estimates. Evaluation of methods reveals that the *window t-test* performs similarly to or better than the best methods [1].

Many biological informations can be used to define gene-sets (metabolic pathway, chromosomal location...). Current methods for gene-set analysis of differential expression are developed to test several hypothesis. We generalize gene-set analysis to answer the main Q_0 question : « Does the individual expression values of the gene-set members differ between two condition ? ». We developed FAERI to answer to this question, by considering 3 criteria : the correlation between genes, the expression level, and the direction of the response (under- or over-expression). *FAERI* is a modified *ANOVA-2* procedure, starting with a two-step reduction of expression data (Z-standardization, directional reduction). *ANOVA-2* is shown to be the best-performing method when analyzing uni-directional gene-sets (all members are either activated, or repressed). *FAERI* reveals to be the most appropriate method for all tested gene-set types.

We developed *PEGASE* to perform differential expression analysis both at the gene and gene-set level. *Consensus* evaluation from several methods is included, to provide users with good results, even if the choice of an optimal method is not easy. Several methods are implemented in *PEGASE*, both at the gene and gene-set level, and performance evaluation can be run based on biological or empirical knowledge. *PEGASE* is also used as a back-end by *PHOENIX*, an online tool for microarray data analysis [2].

1. Berger F., De Hertogh B., Pierre M., Gaigneaux A. & Depiereux E. The "Window t-test": a simple and powerful approach to detect differentially expressed genes in microarray datasets. *Cent. Eur. J. Biol.*, 2008, 3, 327-344.

2. Berger F., De Hertogh B., Bareke E., Pierre M., Gaigneaux A. & Depiereux E. PHOENIX: a web-interface for (re)analyses of microarray data. *Cent. Eur. J. Biol.*, 2009, 4(4) : 603 : 618.

Merci...

... à Bénédicte, mon épouse

Réaliser un doctorat est non seulement un choix scolaire, mais aussi un choix de vie. J'ai entamé ces recherches avant de te connaître, mais tu en as accepté toutes les implications, bien que difficiles à vivre. En particulier, le stage réalisé à Lausanne nous a séparé durant 6 mois, sans parler des nombreuses soirées et nuits passées à rédiger. Merci d'avoir accepté de partager avec moi ces moments difficiles, sans cesser de me soutenir, avec tout ton Amour.

... à Maude, ma nièce
... à Noam, mon neveu

Rédiger une thèse est une tâche difficile, et votre Amour m'a apporté l'énergie et la motivation de mener à bien ce projet. J'arrive au terme d'un parcours académique de haut niveau, mais c'est votre Amour d'enfants qui sont les enseignements les plus importants que j'ai acquis, une force que vous m'avez transmise chacun à votre tour, dès votre naissance.

... à Yves, mon frère
... à Bénédicte, ma belle-soeur
... à Pascal

Quand j'ai choisi de recommencer un cycle universitaire, j'ai passé avec vous énormément de soirées, parfois tardives. Ces soirées en votre compagnie étaient la seule vie sociale que le temps me permettait. Jour après jour, vous n'avez cessé de me soutenir, chacun à votre façon.

... à Marc, mon frère
... à Sabrina

Tout au long de mon parcours universitaire, il a été nécessaire que je me retire au calme pour pouvoir « travailler en paix » et pouvoir avancer sans être dérangé toutes les cinq minutes. Merci à vous deux de m'avoir offert un tel havre de paix, à plusieurs reprises, en toute confiance.

... à mes parents

Vous m'avez tout d'abord donné la possibilité d'entamer des études supérieures. Vous m'avez toujours poussé à croire en mes capacités, et à me perfectionner. J'ai choisi, plus tard, de reprendre des cours. Merci de m'avoir aidé à rendre possible cette reconversion.

... à Nathalie, ma belle-soeur
... à Toni, mon beau frère

Nous nous sommes rencontrés par hasard, et sommes devenus des amis très proches. Nous sommes aujourd'hui de la même famille. En tant qu'ami, vous m'avez traité comme un membre de votre famille. En tant que famille, vous n'avez cessé d'être mes amis. Merci pour tous vos encouragements, et merci

pour toutes les fêtes et soirées organisées ensembles pour décompresser.

... à Brian
... à Henri
... à Wendy
... à Francesca
... à Sylvie

La vie est jalonnée d'allées et venues, certains amis sont de passage, d'autres restent. Vous avez partagé une partie du chemin de la vie à mes cotés, celle de l'amitié. Aujourd'hui, je ne serai pas le même si je n'avais fait votre rencontre.

... à Séba
... à Domino
... à Djids

Nous nous sommes réunis de nombreuses fois pour décompresser ensemble, passer une soirée au coin du feu, écouter de la musique ou méditer, se porter ensemble vers le haut ou s'empêcher mutuellement de se tirer vers le bas, philosopher ou boire un verre... Merci pour vos encouragements, et pour tous ces bons souvenirs.

... à J.A. Martial
... à A. Empain
... à A. Renard
... à B. Joris

Vous m'avez apporté votre appui et vos conseils lorsque j'ai décidé de réorienter ma vie professionnelle et de me consacrer à la bioinformatique, et je tiens à vous en remercier.

... à E. Depiereux, mon promoteur

Merci de m'avoir donné la possibilité de réaliser un doctorat sous ta supervision. Tu as été pour moi non seulement un employeur, mais aussi une personne auprès de qui j'ai pu trouvé conseil, compréhension et respect. Merci également en tant que promoteur, pour m'avoir conduit jusqu'au bout de ce parcours...

... à Benoît
... à Anthoula
... à Eric
... à Bertrand
... à Michaël
... à J.-L. Ruelle
... à S. Gaulis
... à T. Coche
... à M. Remon
... à I. Motte

Vous avez chacun contribué aux discussions et recherches de solutions, et ce document n'aurait pas la même qualité sans votre apport.

... à M. Delorenzi
... à toute l'équipe du BCF
... à D. Goldstein

Merci de m'avoir accueilli aussi chaleureusement au sein de l'institut suisse de bioinformatique, à Lausanne, dans le cadre de ma thèse. Votre disponibilité et nos conversations ont rendu ce stage très constructif.

... à C. Lambert
... à D. Belhomme
... à N. Delsatte
... à F. Wautelet
... à E. Pirotte

Merci pour nos échanges et votre aide lors de l'acquisition, de la configuration, de l'administration et du dépannage du matériel informatique.

... à Monique
... à Olivia
... à Benoît

Vous avez apporté quelques rayons de soleil, beaucoup d'encouragements, et de gaieté, l'intersection entre les amis et les collègues... Merci!

... à la Région Wallonne (DGTRE)
... à GlaxoSmithKline Biologicals
... aux FUNDP

Les recherches présentées dans cet ouvrage ont été financées par un projet FIRST-DEI, sur une convention passée entre la DGTRE, GSK Biologicals, et les FUNDP. Merci d'avoir financé ces recherches par 2 mandats de 2 ans.

... à papy

Je tiens particulièrement à rendre hommage à cet homme qui a quitté son pays, et qui a combattu pour la liberté. Un homme qui a travaillé dans les mines pour nourrir sa famille. Un homme qui a aimé et éduqué celle qui deviendra plus tard ma maman, dans un monde où d'autres laissent leurs enfants payer le prix de leur égoïsme. Merci papy, car tu as été un exemple, un homme de valeurs, de courage, de sacrifice et d'altruisme. Mon intérêt pour les sciences et mon amour de la nature ont tous deux germé dans ton potager...

... et merci à tous ceux qui, un jour, ont croisé ma route !

Table des matières

Avant-propos

J'ai rédigé ce document de façon à permettre une lecture par étapes, de diverses manières. Avant d'entamer l'exposé des recherches, quelques pages décrivent la structure de ce document.

Tout au long de cet ouvrage, des résumés vous permettront d'avoir une vue d'ensemble des principaux chapitres. Ceux-ci sont conçus pour être lus indépendamment des chapitres, et ont été encadrés. De même, chacune des parties importantes comporte une introduction et des conclusions partielles, pour faciliter la lecture, et vous guider lorsque vous parcourrez ces pages.

La majeure partie des résultats présentés ont été publiés dans deux articles. Le premier article présente la méthodologie *window*, mise au point dans le cadre de cette thèse, et a été publié dans la revue *Central European Journal of Biology* [18]. Le second article présente simultanément PEGASE, la méthode *consensus*, et PHOENIX (en collaboration avec Benoît De Hertogh), et a été accepté récemment dans la même revue. La publication commune de PHOENIX et PEGASE est justifiée en raison du duo formé par ces deux outils [19].

Ces quelques remarques introductives étant formulées, je vous souhaite une agréable lecture...

Fabrice Berger

Liste des abréviations

aadUTP : aminoallyl-désoxyribo-uridine tri-phosphate.
ADN : acide désoxyribonucléique.
ADNc : ADN complémentaire.
ADP : adenosine diphosphate.
ANCOVA : analysis of covariance.
ANOVA : analysis of variance. Méthode statistique permettant de déterminer si un traitement ou une combinaison de traitement présente un écart significatif à la moyenne.
ARN : acide ribonucléique.
ARNm : ARN messager.
ARNt : ARN de transfert.
ARNc : ARN complémentaire.
ATP : adenosine triphosphate.
AUC : area under curve.
CDF : channel definition format. Fichier spécifique à un type de microarraus qui spécifie à quel probe sets appartiennent des probes pris individuellement.
Cy3 : fluorophore dont les longueurs d'onde du maximum d'absorption et du maximum d'émission sont de 550nm et 570nm, respectivement.
Cy5 : fluorophore dont les longueurs d'onde du maximum d'absorption et du maximum d'émission sont de 649nm et 670nm, respectivement.
dUTP: 2'-deoxyuridine 5'-triphosphate.
EBAM : Empirical Bayes Analysis of Microarray data.
FAD : flavin adenine dinucleotide (forme oxydée).
FADH2 : flavin adenine dinucleotide plus deux hydrogenes (forme réduite).
SAM : Significance Analysis of Microarray.
FAERI : Functional Analysis : Evaluation of Response Intensities.
FC : fold-change.
FDR : false discovery rate. FP/VN+FP.
FDRoc : courbe ROC modifiée. Sensitivity vs FDR.
FP : Faux Négatifs (erreur de type II). Anglais: FP (false positive).
FP : Faux Positifs (erreur de type I). Anglais: FN (false negative).
FPR : False Positive Rate. FP/VP+FP.
FWER : Family Wise Error Rate.
GO : Gene ontology.
HTML : HyperText Marked Language.
KEGG : Kyoto Encyclopedy of Genes and Genomes.
LPE: *Local Pooled Error*.
MAD : Median Absolute Deviation.
MS: Mean Square. Français: CM (carré moyen).
NAD+ : Nicotinamide Adenine Dinucleotide (forme oxydée).
NADH : Nicotinamide Adenine Dinucleotide (forme réduite).
PEGASE : Performance Evaluation and Global Analysis of Significant Expression.
PCR: Polymerase Chaine Reaction.
PDF : Portable Document Format
PRC: Precision/Recall Curve.
p-value: sous l'hypothèse nulle, la valeur de la probabilité d'obtenir au hasard une valeur

supérieure à la valeur observée.
qRT-PCR : quantitative reverse-transcription PCR
ROC : receiving operator characteristic. Sensitivity vs FPR
Sensitivity: ou puissance = VP/(VP + FN)
SNR ou S2N : Signal to Noise Ratio
Specificity: VN/VN+FP
SSE : *Sum of Squared Error*. Français: SCE: Somme des carrés des écarts
SST : *Total Sum of Squared Error*. Français: SCET: Somme des carrés des écarts totale
SS_A: *Sum of Squared Error (factor A)*. Français: SCE_A: Somme des carrés des écarts factorielle associée au facteur A.
SS_R: *Sum of Squared Error (Residuals)*. Français: SCE_R: Somme des carrés des écarts résiduelle
VN : vrais négatif. Anglais: TN (true negative)
VP : vrais positif. Anglais: TP (true positive)

Vue d'ensemble et organisation des chapitres

Introduction générale

La partie introductive de ce travail vise à définir le contexte dans lequel s'inscrit ce projet de thèse. Nous commencerons par aborder quelques notions de génétique, et par la description succincte des mécanismes impliqués dans la problématique de l'expression des gènes.

La technique de polymérisation en chaîne est ensuite présentée, afin de faciliter la compréhension de cette technique fondamentale pour le lecteur non biologiste.

Enfin, la description expérimentale du contexte s'achève par la présentation de la technologie et de la conception des puces (*one-color* et *dual-color*).

Le second volet de la partie introductive dresse les bases du schéma analytique associé à la technologie, et présente les précautions à prendre lors de la manipulation des données générées. Les principales étapes de prétraitement des données y sont brièvement décrites.

Analyse de l'expression différentielle

Après avoir décrit le contexte dans lequel s'inscrit notre travail, par la définition et la présentation des aspects expérimentaux et technologiques, nous avons brièvement présenté une vue d'ensemble du processus analytique, et les corrections nécessaires des données collectées, grâce aux méthodes de prétraitement.

La seconde partie de ce travail répond au premier objectif visé par le projet: dresser la liste des principales méthodologies décrites et utilisées dans le cadre de l'analyse de l'expression différentielle.

Le chapitre est composé de deux parties, relatives respectivement à l'analyse de l'expression différentielle par gène, et par groupe de gènes.

Dans chacune des deux parties, un résumé introductif présente globalement les différents types de méthodes existantes, afin d'en rassembler les informations principales, avant de présenter un échantillon de méthodes représentatives des différentes procédures statistiques employées.

Résultats

Le troisième chapitre de cet ouvrage présente les résultats des recherches menées sur la

thématique de l'expression différentielle.

Pour répondre aux objectifs visés par le travail, les travaux réalisés sont présentés en trois parties.

Au sein de la première partie, l'objectif général visé est le partage d'information entre les gènes pour en améliorer l'estimation de la variance individuelle. Nous répondrons à la problématique de l'analyse individuelle (relative aux gènes) en présentant tout d'abord une caractérisation de la structure des données, en regard de la relation observée empiriquement entre le niveau d'expression et la variabilité individuelle. Nous en tirerons des conclusions sur son usage, avant de matérialiser la stratégie par le développement d'une méthode originale, le *window t-test*. Celle-ci sera ensuite comparée à d'autres méthodes pour en dégager des similitudes et enseignements théoriques, avant d'être évaluée en terme de performances. Les évaluations présentées ont été réalisée au départ de trois types de jeux de données (simulés, *spike-in* et biologique), et montreront au lecteur le gain de performances lié à la démarche proposée. Ces tests nous ont permis de publier le *window t-test* dans la revue *Central European Journal of Biology* [18].

Cette première partie s'achève avec la présentation d'une autre approche originale, basée sur le *consensus* de plusieurs méthodes, visant à obtenir un résultat plus robuste. Les évaluations réalisées montreront que la méthode permet d'obtenir des résultats associés à des performances similaires aux meilleures méthodes, bien que l'on se soit rendu compte par ailleurs que la méthode la plus performante dépend du jeu de données, du nombre de réplicats… et qu'aucune ne peut prétendre être la panacée. L'utilisation de la méthode *consensus* permet à l'expérimentateur « naïf » de ne pas devoir se soucier du choix d'une méthode particulière, au risque de se cantonner à une méthode par défaut qui n'est peut être pas optimale dans son contexte.

La seconde partie des résultats présentés s'intéresse à l'analyse des données d'expression en groupant les gènes sur base de critères connus (à ne pas confondre avec l'étude de la coexpression des gènes). Après un comparatif des méthodes disponibles qui récapitule les avantages et inconvénients de chaque méthodes, nous proposons de répondre au problème sur base des enseignements apportés par la procédure *ANOVA-2*. Plusieurs ajustements de la méthode *ANOVA-2* sont présentés, et conduisent à la définition d'une nouvelle méthode d'analyse de groupe, dénommée *FAERI*, pour permettre l'étude des groupes dont les membres présentent une réponse mixte, en sur-expression et en sous-expression (groupes bidirectionnels).

Les performances de ces deux stratégies analytiques sont ensuite évaluées d'une part sur

base de simulations, d'autre part en prenant un exemple d'analyse de données réelles.

Les simulations ont pour objectif de quantifier l'impact de différents critères mentionnés dans la littérature scientifique (la taille des groupes, la corrélation entre les membres, la proportion de gènes impliqués, la direction de la réponse), et montreront que la procédure *FAERI* offre des performances supérieures à toutes les autres méthodes testées, pour chaque cas de figure. Seule exception à cette conclusion: la méthode *ANOVA-2* (inutilisée actuellement) effectue plus efficacement l'analyse de groupes au sein desquels tous les membres présentent une réponse dans la même direction (en sur-expression, par exemple).

L'analyse d'un jeu de données réelles révélera ensuite que cet effet est visible dans les résultats: les groupes détectés par l'ensemble des méthodes se réfèrent aux mêmes mécanismes cellulaires et moléculaires, mais les méthodes unidirectionnelles et bidirectionnelles fournissent des résultats différents, principalement liés à la nature même des groupes étudiés. De plus, la comparaison des résultats entre plusieurs jeux de données relatifs au même thème attribue la plus forte corrélation à *FAERI*, quelle que soit la source de définition des groupes.

La troisième partie des résultats répond à la démarche d'intégration de notre expérience dans un logiciel automatisé, *PEGASE*, qui intègre l'ensemble des démarches suivies tout au long de nos recherches. Nous y présentons tout d'abord la stratégie générale que nous proposons de suivre, sur base des enseignements issus des deux premières parties. Nous présentons ensuite rapidement les critères qui ont été pris en compte pour traduire la stratégie analytique au sein de la structure de *PEGASE*, avant de présenter un tour d'horizon des principales étapes de l'analyse et des méthodes implémentées pour y répondre.

En particulier, l'évaluation des performances, qui représente une part importante des investigations menées, y est discutée en regard de notre expérience et de nos conclusions.

Cette troisième partie s'achève avec la présentation de *PHOENIX*, un outil disponible en ligne, qui repose sur *PEGASE* pour effectuer les analyses. *PHOENIX* et *PEGASE* ont également été acceptés dans la revue *Central European Journal of Biology*, dans un article qui présente également la méthode *consensus* (présenté dans la première partie) [19].

A titre d'exemple complémentaire de l'utilisation de *PEGASE*, nous présenterons une analyse réalisée sur un jeu de données simulé au départ de données réelles. Ce jeux a été

mis au point par Benoît De Hertogh et Bertrand De Meulder, sur base des idées issues de la rencontre entre nos recherches. Pour répondre au limitations actuelles des procédures de *benchmark*, plusieurs jeux de données réelles y sont utilisés.

Nous tirerons ensuite des conclusions générales quant à l'analyse de l'expression de groupes de gènes, à la lumière des résultats obtenus, pour en dégager des perspectives d'extension des différentes stratégies envisagées, et la définition de projets originaux.

Le dernier chapitre du travail présenté ici rassemble les descriptions procédurales, et la présentation des jeux de données et logiciels utilisés pour mener à bien nos recherches.

I. Introduction Générale

I.A. Introduction

Les avancées réalisées dans le domaine de la biologie moléculaire ont conduit, ces dernières années, à l'utilisation de technologies diverses fournissant une grande quantité d'informations. En particulier, l'ère post-génomique se voit caractérisée par l'apparition de techniques expérimentales permettant de réaliser des mesures à l'échelle du génome. Parmi ces techniques s'inscrivent les biopuces, couramment dénommées *microarrays* ou *microchips*, permettant d'étudier le niveau d'expression associé à chacun des gènes [5, 37, 81, 110].

Dès lors, nous sommes aujourd'hui en mesure (du point de vue expérimental) de comparer les profils d'expression à l'échelle du génome dans différentes conditions ou pour différentes pathologies. A titre d'exemple, des avancées non négligeables ont été réalisées en oncologie, conduisant à une meilleure caractérisation des perturbations biochimiques qui interviennent dans différents types de cancers. Sur base des profils d'expression, plusieurs études ont en outre démontré que le recours à cette technologie permet de diagnostiquer différents types de cancers, grâce à une méthode de classification automatique [65].

Cependant, bien que leur utilisation soit de plus en plus répandue, le coût de ces expériences constitue toujours un obstacle à leur utilisation routinière, limitant le nombre de mesures réalisées. En conséquence, les méthodes classiques d'analyse statistique des données doivent être adaptées, pour relever le défi que pose le nombre restreint de mesures réalisées sur un nombre important d'échantillons différents. Classiquement, le niveau d'expression de plusieurs milliers de gènes (voir plusieurs dizaines de milliers) est évalué à partir de 2, 3, 4 ou 5 réplicats. L'analyse de tels jeux de données, compte tenu tant de leur importance biologique que des compromis nécessaires entre la complexité des analyses réalisées et les approximations dictées par le temps de calcul disponible constitue aujourd'hui l'un des enjeux majeurs des recherches en bioinformatique.

Pour être correctement interprétées, les données collectées grâces aux puces à ADN doivent subir une série de transformations. L'intensité de fluorescence sera traduite en données chiffrées normalisées. La comparaison des valeurs d'expression génique sera ensuite réalisée pour différentes conditions. A l'issue de cette seconde étape, une probabilité d'être différentiellement exprimé sera associée à chaque gène. Les gènes mis en évidence sont triés sur base de leur profil d'expression dans différentes conditions. L'annotation automatique des gènes à l'aide de bases de données génomiques/biochimiques permet alors d'interpréter les résultats de l'expérience à la lumière des connaissances actuelles. Ainsi, par exemple, la base de données *GO* (Gene Ontology) permet d'associer à chaque gène des informations sur sa fonction, la localisation de la protéine codée, les gènes éventuellement régulés (dans le cas d'un facteur de transcription), ... La base de données *KEGG* permet quant à elle de replacer le gène étudié dans la(les) voie(s) biochimique(s) au sein de laquelle (desquelles) il s'inscrit [5, 37, 81, 110].

A chacune de ces étapes sont associées des méthodes d'efficacité variable conduisant à des erreurs systématiques. La biostatistique s'attache à minimiser ces erreurs, et à adapter les traitements statistiques réalisés de façon à retrouver l'information biologique portée par les *microarrays*, laquelle peut-être masquée par le bruit inhérent à la technologie, par les erreurs issues des approximations réalisées et par le petit nombre de réplicats. Le travail présenté ici se place en aval du prétraitement des données pour se concentrer sur les méthodes statistiques de détection des gènes significativement sur- ou sous-exprimés.

I.B.
Description expérimentale

I.B.1. Le décodage de l'ADN

L'ADN contenu dans le noyau de nos cellules peut être comparé à un livre de recettes, qui porte les instructions nécessaires au bon fonctionnement cellulaire. Cette information est codée grâce à la succession des désoxyribonucléotides A, C, G, T (Adénosine, Cytidine, Guanosine, Thymidine). Selon les besoins cellulaires, l'activation d'un gène conduit à son expression. La transcription de l'ADN en ARNm (ARN messager) crée une copie « utilisable » de cette information, sur base de la complémentarité des nucléotides (A et T sont complémentaires, de même que C et G) [94].

L'ARN, contrairement à l'ADN, est constitué des quatre nucléotides A, C, G et U (Uridine), pour lesquels chaque base est fixée à une molécule de ribose (par opposition au désoxyribose de l'ADN). L'uridine est un nucléotide analogue à la thymidine, et s'apparie avec l'adénosine.

L'ADN double brin est transcrit en une molécule d'ARN qui porte l'information nécessaire pour synthétiser une protéine, le cas échéant. La traduction est réalisée par les ribosomes, qui utilisent ARNt (ARN de transfert) par complémentarité avec l'ARNm pour synthétiser une chaîne polypeptidique. L'ARNt est associé à un acide aminé, spécifiquement, et porte une séquence de reconnaissance d'un motif de trois lettres, par complémentarité avec la séquence de l'ARNm. Chaque combinaison de trois nucléotides définit un codon. Le code génétique est une table de correspondance entre ces motifs et les acides aminés associés. Les 64 codons déterminés par toutes les combinaisons possibles de trois nucléotides sont associés à 20 acides aminés. Le code est dit « dégénéré » car plusieurs codons peuvent coder pour le même acide aminé. De plus, quatre codons sont associés à un rôle régulateur important : le codon « *start* » AUG, qui code pour la méthionine et les codons « *stop* » UAA, UAG et UGA, également appelés « ambre », « ocre » et « opale ». L'ensemble des codons compris entre le codon « *start* » et l'un des codons « *stop* » détermine une séquence en acides aminés (polypeptide), qui forme une protéine (ou une sous-unité protéique).

Chaque molécule d'ARN porte des instructions, mais celle-ci ne conduit pas systématiquement à l'expression d'une protéine. Les ARNt (ARN de transfert) sont utilisés en combinaison avec les acides aminés et les ribosomes, et assurent le lien entre l'information génétique et les protéines produites, grâce au code génétique. L'ARNr (ARN ribosomal) est un constituant des ribosomes. Les molécules d'ARN peuvent également présenter une activité régulatrice (par exemple, la séquence complémentaire d'un autre

ARN conduit à une hybridation qui perturbe la traduction), ou une activité enzymatique (les ribozymes).

L'organisation du génome des organismes eucaryotes est plus complexe. D'une part, celui-ci est fragmenté en plusieurs chromosomes. Chaque chromosome porte les séquences codantes des différents gènes, ainsi que plusieurs motifs utilisés pour la régulation de leur activité (éléments « cis », proches, et « trans », distants, entre autres). Les séquences génétiques sont de plus « morcelées », la séquence propre à un gène étant interrompue régulièrement par des séquences non codantes, dénommées « introns ». Les régions qui portent l'information génétique sont appelées « exons ». Pour pouvoir exploiter l'information génétique, l'ARN, transcrit au départ de l'ADN, doit être « nettoyé », grâce à un processus appelé « épissage ». Au terme de cette étape, l'ARN messager est produit (ARNm).

Au départ d'un seul ARN transcrit, des éléments de régulation peuvent influencer le processus d'épissage, de sorte que plusieurs ARN messagers puissent être obtenus, par épissage alternatif, au départ d'une seule séquence.

En conséquence, l'étude de l'expression des gènes, quel qu'en soit le produit, repose sur l'ARNm obtenu à l'issue de l'épissage. L'ARN, de part ses propriétés biophysiques, est instable, à l'inverse de l'ADN. L'étude de l'ARN nécessite donc de retranscrire celui-ci en ADN double brin par une opération appelée transcription inverse. Ceci peut être réalisé *in vitro*, grâce à l'utilisation d'enzymes appelées « *reverse transcriptases* ». Ces enzymes ont été découvertes dans les virus dont le génome est constitué d'ARN, et dont la survie est assurée par la copie du génome sous forme d'ADN double brin, et par l'insertion de celui-ci dans le génome des cellules infectées.

Il est important de mentionner également que la régulation de l'activité cellulaire ne repose pas uniquement sur la transcription et la traduction de l'ADN. La présence d'un ARNm ne signifie pas pour autant que celui-ci soit traduit en protéine. Au même titre, la production d'une protéine n'implique pas que celle-ci soit active. Plusieurs modifications post-traductionnelles sont souvent nécessaires à son activation (phosphorylation, glycolisation...). De plus, l'activité d'une protéine peut être dépendante de la présence d'un partenaire (une autre protéine, un cofacteur, un enzyme ou coenzyme, réactif...). L'étude de l'expression des gènes via la détection de l'ARN produit n'implique donc pas nécessairement que l'ensemble des ARN détectés soient impliqués dans l'expérience réalisée.

I.B.2. Les réactions de polymérisation en chaîne (PCR)

Pour étudier individuellement l'expression des gènes, plusieurs techniques de biologie moléculaire et de génie génétique ont été mises au point. Parmi celles-ci, l'une des plus importantes est la PCR. Lors du cycle cellulaire, une phase de synthèse permet de dupliquer l'ADN avant la division cellulaire, pour assurer la transmission du matériel génétique entre la cellule mère et les cellules filles. L'ADN, double brin, est tout d'abord clivé, pour permettre aux polymérases de s'attacher à chacun des brins séparés. Les polymérases utilisent ensuite l'ADN simple brin pour le convertir en ADN double-brin, grâce à la complémentarité des nucléotides [141].

La *PCR* repose sur l'utilisation *in vitro* de l'activité des polymérases pour recopier une séquence d'ADN au cours de plusieurs cycles, afin de la produire en grande quantité. Le produit de la réaction est ainsi plus facile à détecter et à purifier pour une réutilisation ultérieure.

De plus, la *PCR* utilise des séquences de reconnaissance spécifiques, utilisées comme amorces, pour amplifier uniquement le contenu ciblé (l'amplicon). La figure I.B.1 en illustre le principe.

La polymérisation en chaîne consiste à séparer physiquement les deux brins de l'ADN grâce à une élévation de température qui induit sa dénaturation. Ensuite, la température est réduite jusqu'à une température d'hybridation des amorces. Celle-ci détermine le succès de la réaction et de sa spécificité. Le choix de la longueur et de la séquence des amorces détermine leur spécificité, et leur composition nucléotidique détermine la température à laquelle les amorces peuvent être hybridées à l'ADN. Une température trop basse présente le risque d'hybridations non spécifiques.

Deux amorces sont utilisées, chacune étant spécifique de l'une des extrémités de la séquence ciblée. Une fois cette hybridation réalisée, la température est à nouveau diminuée, pour permettre aux polymérases de se fixer aux régions double brin formées par l'hybridation des amorces. Les polymérases progressent alors le long de l'ADN en ajoutant les nucléotides suivants par complémentarité. La durée du cycle dépend de la longueur de la séquence ciblée, et de la vitesse de progression des polymérases. Au terme du premier cycle, chacune des amorces a été utilisée pour dupliquer la séquence ciblée [141].

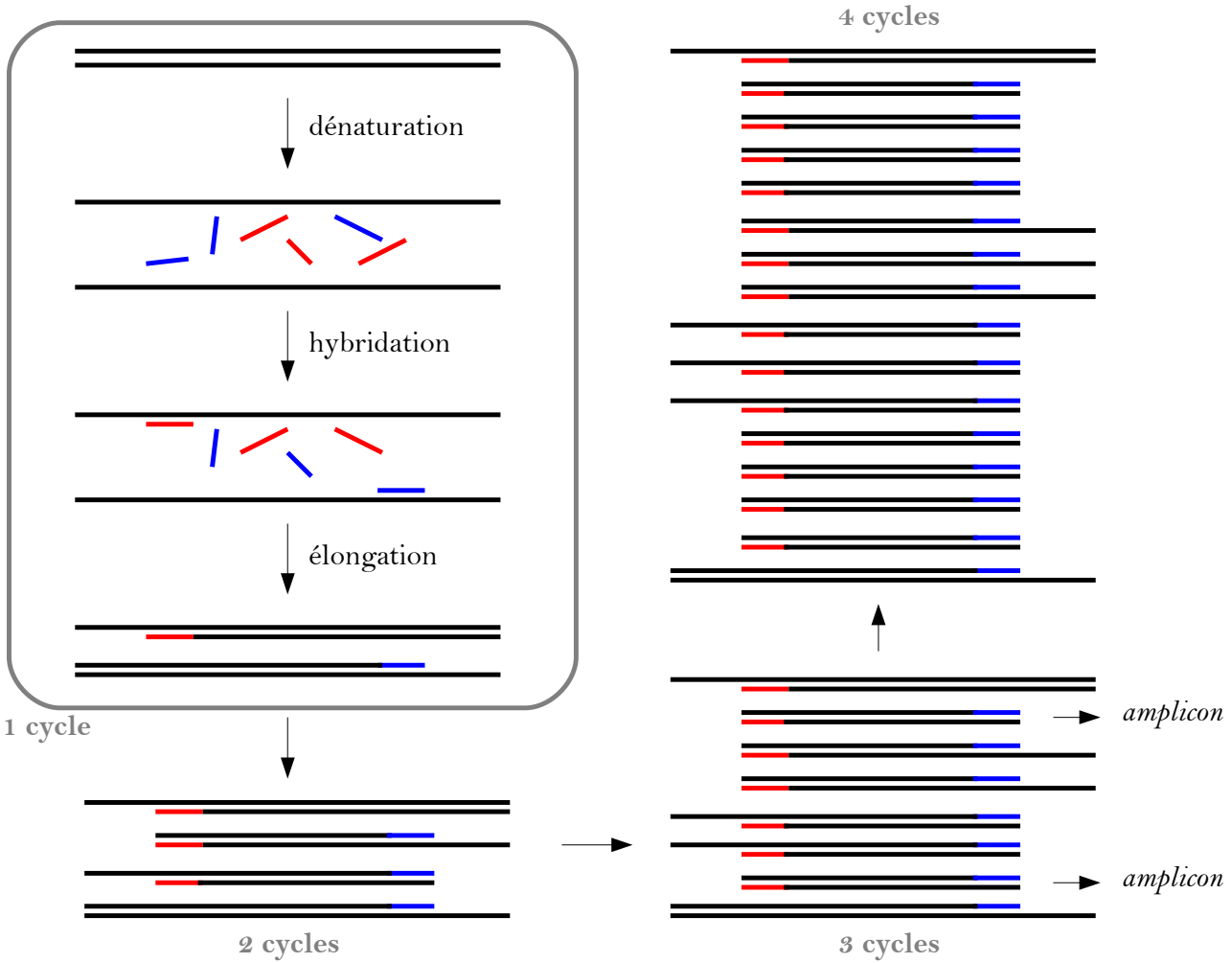


Figure I.B.1 : Illustration du principe de la réaction de polymérisation en chaîne (PCR). Chaque cycle repose sur la séparation des brins d'ADN, suivie par une hybridation des amorces, qui sont ensuite utilisés pour recopier les brins d'ADN. Au bout de 3 cycles, 2 amplicons sont créés (pour 2^3 molécules d'ADN). Au bout de 4 cycles, 8 amplicons sont disponibles, pour 2^4 molécules d'ADN.

Le processus est ensuite répété, enchaînant une température de dénaturation, une température d'hybridation des amorces, et une température d'élongation de la chaîne double brin formée par les amorces et le matériel ciblé. A chaque cycle, l'information est dupliquée, et le nombre de cycles détermine la quantité finale d'ADN amplifié. Mathématiquement, après n cycles réalisés sur une seule copie d'ADN, $2^n - 1$ copies sont ainsi générées [141].

Grâce à cette technologie, des avancées considérables ont pu être réalisées dans le domaine de la biologie moléculaire. Plusieurs techniques dérivées de ce principe général permettent de déterminer si un gène précis est présent dans l'échantillon analysé, et d'en déterminer la concentration (*qRT-PCR*). La technique est également utilisée couramment pour isoler un gène, et le modifier grâce à des amorces prévues à cet effet pour créer un ou plusieurs variants du gène, en apportant une correction déterminée (amorces modifiées) ou en introduisant des mutations au hasard (évolution dirigée).

La *PCR* peut être utilisée en combinaison avec une procédure qui fait intervenir des transcriptases inverses, lorsque la séquence ciblée est portée par des molécules d'ARN, afin d'en faire une copie sous forme d'ADNc (ADN complémentaire).

Au terme du processus, le gène amplifié peut être purifié et utilisé par d'autres techniques complémentaires, ou réintroduit dans des cellules vivantes. A titre d'exemple, la *PCR* est régulièrement utilisée pour créer des protéines-fusions, ou le choix des amorces permet de mettre bout à bout les séquences qui codent pour une protéine d'intérêt et une protéine fluorescente. La fluorescence de la protéine-fusion permet alors de visualiser la production de la protéine dans différents tissus ou organes, à différentes phases du développement... Il est également possible de placer l'expression d'une protéine fluorescente sous le contrôle d'une séquence de régulation choisie, afin d'en étudier le comportement [141].

I.B.3. Les puces à ADN

L'étude systématique de l'expression de l'ensemble des gènes présents au sein d'un génome représente un travail d'une ampleur considérable. Plusieurs technologies, dites « à haut débit », ont été développées récemment afin d'étudier l'ensemble du génome de façon automatique. Parmi elles, des puces ont été développées pour sonder spécifiquement chacun des gènes, sur base de procédures robotisées, pour offrir une vue d'ensemble du profil d'expression. Les techniques utilisées traditionnellement en biologie moléculaire permettent ensuite de valider une partie de ces résultats, et d'étudier plus en profondeur certains mécanismes.

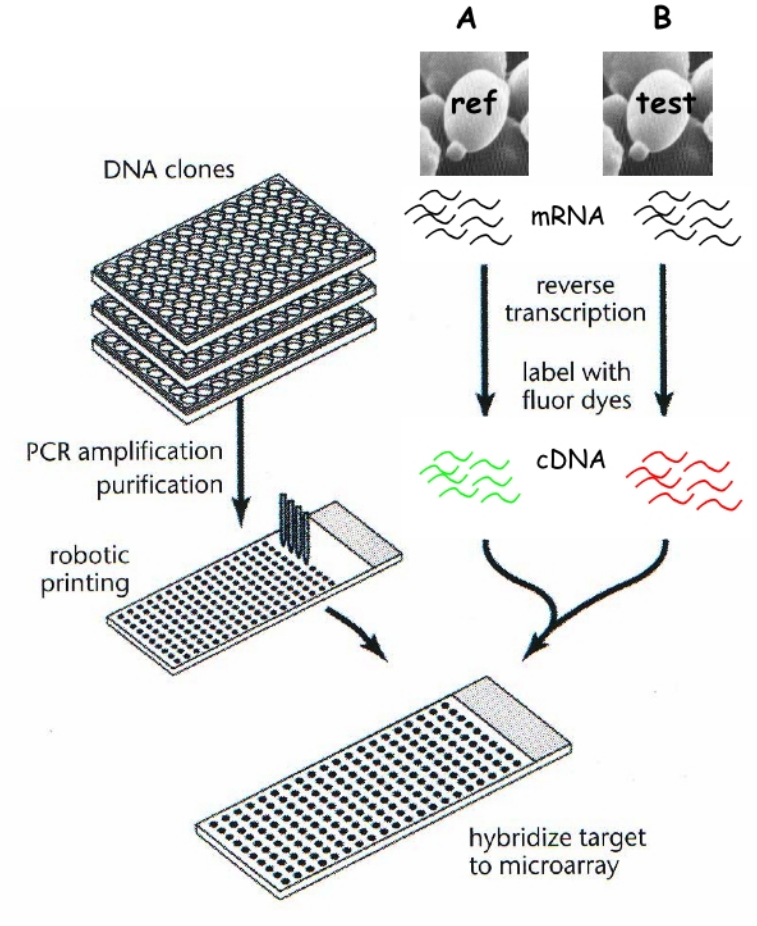
L'étude de l'expression des gènes effectuée par ce biais permet d'envisager des avancées considérables tant en recherche fondamentale qu'en recherche appliquée ou clinique. Ainsi, la comparaison du profil d'expression de tissus sains et de tissus contaminés permet de dégager des informations sur les mécanismes biologiques associés à une pathologie. La comparaison d'échantillons associés à divers médicaments permet d'en comprendre les effets, et de mettre au point de nouveaux traitements thérapeutiques. Enfin, la comparaison du profil d'un échantillon avec des profils d'expression préalablement décrits permet de dégager des similitudes, voire de diagnostiquer une pathologie.

Les puces à ADN, aussi appelées *microarrays* ou *microchips*, reposent actuellement sur deux techniques différentes de préparation des puces (*spotted arrays* et *oligonucleotide arrays*), et deux modes de marquage fluorescents (*one-color* et *dual-color*).

I.B.3.a. « Spotted arrays »

La conception des *spotted arrays* est illustrée dans la figure I.B.2 (schéma de gauche). La puce repose une première étape au cours de laquelle la synthèse des *probes* est réalisée. Plusieurs approches différentes sont envisagées au cours de cette étape de synthèse. Les *probes* peuvent être des oligonucléotides définis par spécificité vis-à-vis des différents gènes, des produits d'amplification par *PCR* ou de l'ADN complémentaire (ADNc).

Une fois purifiées, un système robotisé permets de les déposer sur une plaque, identifiés par leur position. La plaque ainsi préparée est prête pour l'hybridation des échantillons.



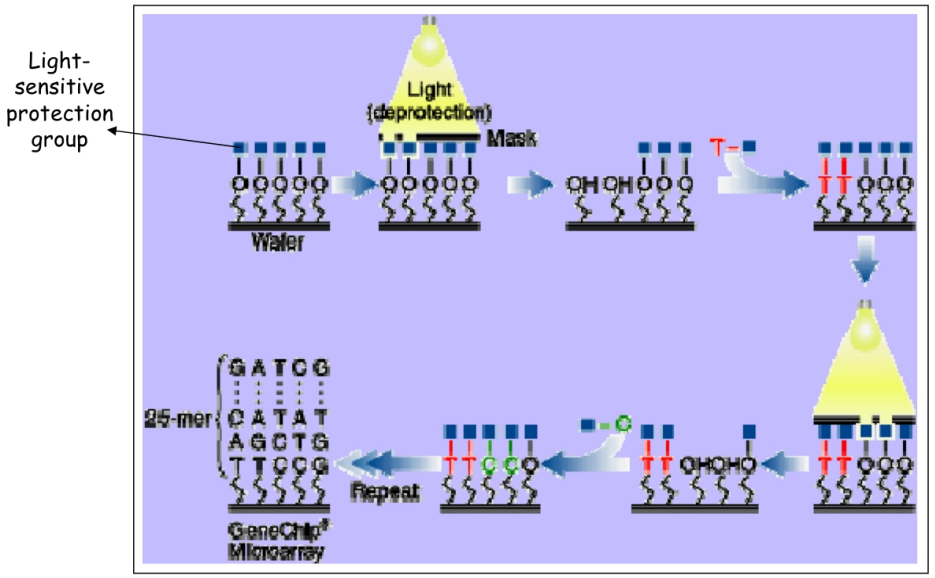


Figure I.B.2 : Illustration de la stratégie de conception des *spotted arrays* (à gauche) et des *oligonucleotide arrays* (à droite).
Le schéma de gauche montre que les *spotted arrays* reposent sur deux étapes : la synthèse des probes, et le dépôt de celles-ci sur une plaque.
Le schéma de droite illustre un autre principe de conception des puces, par synthèse des oligonucléotides directement sur la puce, grâce à la photolithographie. Ce principe est utilisé au sein des puces Affymetrix de type *GeneChips*.
Source: Bruno André, cours donné dans le cadre du DEA Inter-Universitaire en Bioinformatique, 2004-05.

I.B.3.b. « Oligonucleotide arrays »

La procédure utilisée pour construire la biopuce repose sur la photolithographie, la chimie, et l'utilisation de semi-conducteurs. Les *probes* sont directement construits sur la puce, nucléotide par nucléotide, tel qu'illustré dans la figure I.B.2.

La procédure commence par la fixation d'une molécule sensible à la lumière. Un masque est utilisé, et l'illumination de la plaque à travers ce masque détruit la molécule photosensible. A l'issue de cette première opération, un groupe hydroxyle (OH) est disponible à chaque endroit éclairé. La plaque est ensuite mise en présence du premier nucléotide, dont le groupement phosphate (extrémité 5') forme une liaison avec le groupement OH libéré.Pour reproduire l'opération, l'extrémité 3'-OH du nucléotide ajouté est également protégée par une molécule photosensible. A l'issue de la première étape, la totalité de la puce est à nouveau protégée, et un nucléotide est fixé à certaines positions, qui correspondent à la localisation des *probes* qui commencent par ce nucléotide. La procédure est reproduite successivement avec les 4 nucléotides, en utilisant un masque différent à chaque étape de la synthèse.

I.B.3.c. Principe de la technologie « one-color »

Le terme *one-color*, ou *one-dye*, se réfère à l'usage d'un seul fluorophore lors du marquage des échantillons. La méthode exploite l'affinité de la biotine et de la streptavidine.

Dans une première étape, l'ARN total est extrait. L'ADNc est obtenu grâce à une transcription inverse réalisée *in vitro*. L'ADNc produit est ensuite transcrit *in vitro* en ARNc, simple brin. L'un des nucléotides utilisés au cours de la synthèse est le dUTP, sur lequel est fixé la biotine. Celui-ci est incorporé à l'ARNc en formation.

L'ARNc obtenus, est ensuite fragmenté, et les fragments d'ARNc biotinilés sont hybridés sur la biopuce. Après une étape de nettoyage, la streptavidine est utilisée sur la biopuce, et par affinité avec la biotine. L'intensité de fluorescence mesurée est liée à la quantité de biotine présente, elle-même liée à la quantité d'ARNc hybridé.

Chaque échantillon est hybridé sur une biopuce différente, et la comparaison des profils d'expression repose sur la comparaison de puces différentes.

I.B.3.d. Principe de la technologie « dual-color »

Le principe général des puces à ADN faisant usage de deux couleurs est illustré dans la figure I.B.3.

L'ARN messager des échantillons que le biologiste moléculaire souhaite analyser est tout d'abord extrait, et utilisé pour produire de l'ADN complémentaire, grâce à la transcriptase inverse (*reverse transcriptase*). Au cours de cette étape, nécessitant la présence des quatre nucléotides, un cinquième nucléotide est ajouté : l'aminoallyl-désoxyribo-uridine tri-phosphate (aadUTP). Il s'agit d'un analogue de la thymidine, qui est intégré à l'ADNc en formation par complémentarité avec l'adénosine de l'ARN messager.

L'ARN messager est ensuite dégradé au cours de l'étape de nettoyage, et l'ADNc simple brin est conservé.

La seconde étape repose sur le marquage de l'ADNc, au cours de l'étape de *labeling*. L'ADNc relatif à l'une des deux conditions comparées est marqué, grâce à l'aadUTP, avec le fluorophore Cy3, dont les longueurs d'onde du maximum d'absorption et du maximum d'émission sont de 550nm et 570nm, respectivement. L'échantillon relatif à la seconde condition est quant à lui marqué sur le même principe avec le fluorophore Cy5, dont les pics d'absorption et d'émission sont situés à 649nm et 670nm, respectivement.

Enfin, les deux échantillons, marqués séparément, sont hybridés ensemble sur la microplaque, et les mesures d'intensité relatives aux deux conditions sont effectuées séparément en utilisant deux lasers dont la longueur d'onde correspond au maximum d'absorption de Cy3 et de Cy5. Les mesures d'intensité sont prises à la longueur d'onde qui correspond au maximum d'émission des deux fluorophores.

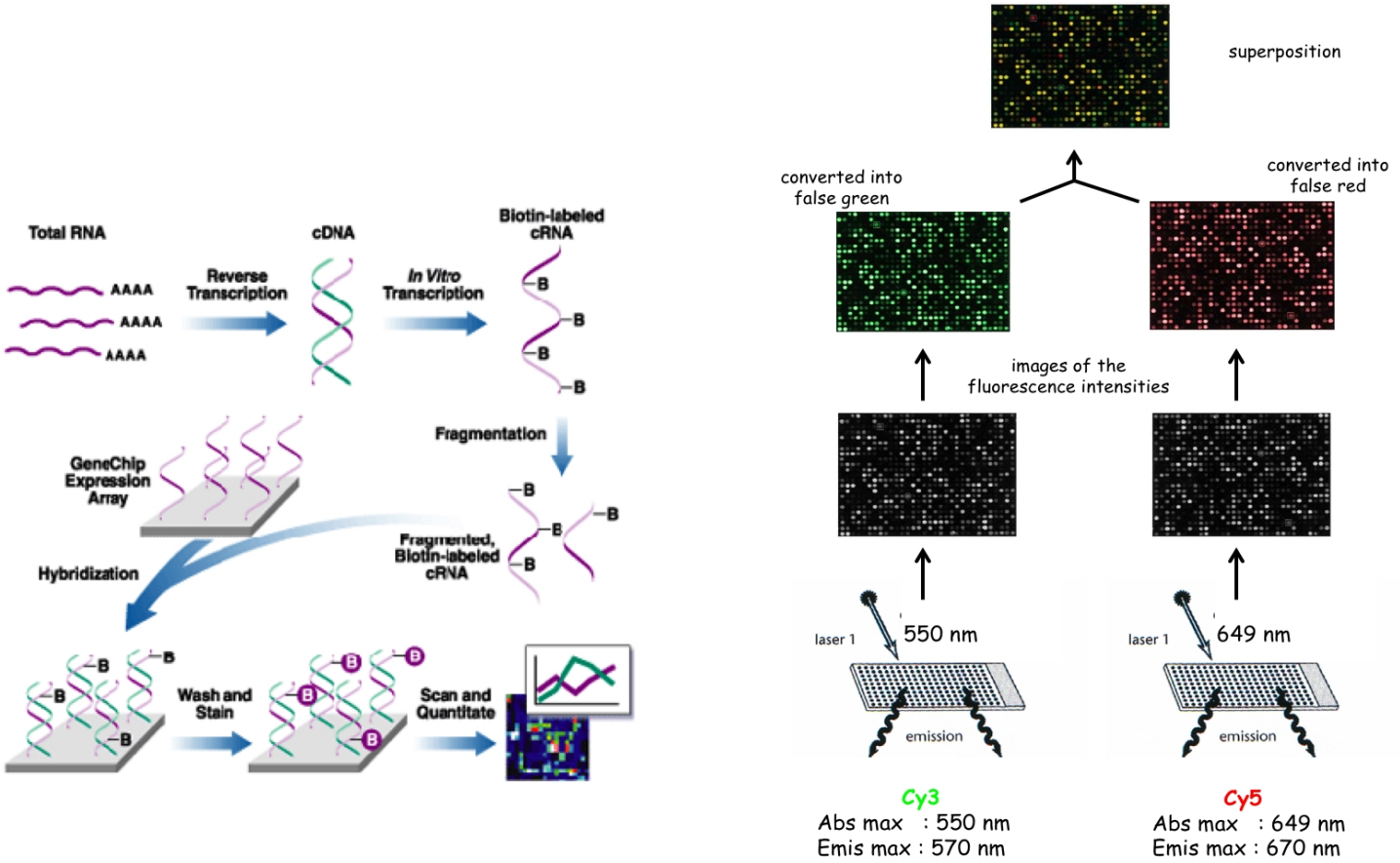


Figure I.B.3 : Illustration des deux types de marquage associés à la technologie des puces à ADN, *one-color* (à gauche) et *dual-color* (à droite).
Le schéma de gauche correspond au mode de marquage utilisé au sein des puces Affymetrix *GeneChip*, grâce à l'affinité de la biotine et de la streptavidine. Chaque échantillon comparé est hybridé sur une puce différente. A l'inverse, dans le cas du marquage en deux couleurs, les fluorophores Cy3 et Cy5 sont utilisés pour marquer deux échantillons différents, qui sont ensuite hybridés sur la même puce. Les mesures propres aux deux échantillons sont réalisées aux deux longueurs d'ondes d'émission des fluorophores.
Source: Bruno André, cours donné dans le cadre du DEA Inter-Universitaire en Bioinformatique, 2004-05.

Les mesures associées à l'échantillon marqué au Cy3 sont représentées en vert. Celles relatives au marqueur Cy5 sont représentées en rouge. Les deux couleurs sont superposées sur une même photographie, et la couleur résultante est représentative de la différence d'expression du gène concerné, pour chacun des gènes représentés sur la microplaque.

Etant donné que les deux échantillons sont hybridés sur la même plaque, les jeux de données issus de la technologie deux-couleurs fournissent des résultats pairés.

I.B.3.e. Les puces Affymetrix « GeneChips »

Les puces à ADN de la technologie *GeneChips*, produites par la société Affymetrix, sont très répandues. Les biopuces de ce type reposent sur le principe des *oligonucleotides arrays*, et sont conçus par photolithographie. Le marquage des échantillons repose sur le modèle *one-color* et sur la fluorescence de la biotine en présence de streptavidine.

Les micropuces sont réalisées en utilisant de courtes séquences d'ADN, dénommées *probes*. Chaque *probe* est un oligonucléotide d'une longueur de 25 nucléotides. Chaque gène est représenté par au moins 11 *probes*, réparties uniformément sur la puce. Chacune des *probes* est définie en recherchant les séquences spécifiques de chaque gène dans une région de 600 paires de bases, située à l'extrémité 3' du gène. L'ensemble des *probes* relatifs à un même gène constituent un *probeset*.

Un second jeu de *probes* est également défini en introduisant des erreurs au niveau du nucléotide central des *probes* définis, pour pouvoir évaluer la qualité de l'hybridation par comparaison entre les mesures effectuées sur les *probes* spécifiques (*perfect match* et *mismatch*).

Chaque échantillon est marqué avec un même fluorophore, et est hybridé sur des puces différentes. La comparaison des profils d'expression repose donc sur la comparaison de l'intensité lumineuse observée sur les différentes puces.

I.C.
Description de l'analyse

I.C.1. Présentation générale

Avant d'utiliser les mesures d'intensité obtenues sur *microarrays*, celles-ci doivent tout d'abord être normalisées, sur base de différents critères, afin d'être comparables entre elles et donc analysables globalement. Idéalement, chaque source de variabilité doit être prise en compte. Leur caractérisation étant incomplète ou inexistante, les outils actuels de normalisation sont limités à certains critères contrôlables, tels que le nombre de cellules présentes dans l'échantillon, l'efficacité de l'extraction de l'ARN total, de l'isolation de l'ARN messager, du marquage utilisé (*labeling*), de l'hybridation et de la qualité de la mesure du signal. Ainsi, les méthodes de prétraitement actuelles reposent sur plusieurs étapes de normalisation des données, respectivement liées à la quantité totale d'ARN (ou d'ARN ribosomal), à l'expression normale des gènes (*housekeeping genes*), à de l'ARN de référence, et à une mise à la même échelle des données finales [67].

L'expression des gènes peut être étudiée à trois niveaux, de complexité croissante.

Premièrement, chaque gène peut être étudié séparément, grâce aux méthodes d'analyse individuelle. L'objectif de ces études est l'identification des gènes impliqués dans la différence de réponse entre les conditions comparées.

Le second niveau d'étude est basé sur la coexpression de plusieurs gènes. Deux stratégies d'analyse sont envisageables. Dans le premier scénario, l'analyse de groupes, l'expression de plusieurs gènes biologiquement liés par des critères connus peut être étudiée, afin de mettre en évidence l'implication de fonctions communes à plusieurs gènes, de mécanismes de régulation communs ... Le second scénario d'étude de la coexpression des gènes, appelée *group discovery*, intervient lorsque les informations biologiques disponibles ne permettent pas de guider l'analyse, et vise à découvrir des mécanismes régulateurs ou fonctions communes sur base de l'observation de la coexpression de gènes au départ des données expérimentales (méthodes de *clustering*).

Enfin, le troisième niveau d'étude des profils d'expression vise à comprendre les connexions entre les différents gènes et groupes de gènes, voies métaboliques, et de modéliser les réseaux de gènes et de protéines responsables des profils d'expression observés [67].

I.C.2. Les sources de variabilité biologiques et techniques

La principale source d'erreur rencontrée lors de l'étude de l'expression différentielle sur base de données issues de *microarrays* est la variabilité des résultats obtenus. Cette variabilité est influencée par de nombreux facteurs, dont l'impact individuel reste peu décrit, ce qui rend difficile la modélisation de ces facteurs de variabilité. Globalement, ils peuvent être classés en deux catégories : les facteurs biologiques et les facteurs expérimentaux.

La première et principale catégorie de sources de variabilité est attribuée aux facteurs biologiques, qui sont de différentes natures. Des cellules génétiquement identiques, cultivées et traitées identiquement présentent des profils d'expression variables, qui correspondent à un contexte biologique différent entre les cellules :

- ☞ l'environnement local peut être différent (gradients de température, de nutriments...) ;
- ☞ la phase de croissance cellulaire peut être différente ;
- ☞ le cycle cellulaire peut être décalé ;
- ☞ la distribution du contenu cellulaire lors de la division peut être inégale ;
- ☞ l'expression de certains gènes peut dépendre uniquement de quelques molécules ;
- ☞ l'expression des gènes peut changer très vite.

Tous ces paramètres influent sur la régulation de la transcription, ainsi que sur le niveau d'expression atteint [67].

La seconde catégorie de sources de variabilité est liée aux procédures expérimentales utilisées. Chaque étape du traitement effectué sur les échantillons est une source supplémentaire de variabilité (culture cellulaire ou échantillon prélevé, méthodes d'extraction, d'amplification, d'isolation, conditions d'hybridation, contamination par d'autres acides nucléiques...). Une attention particulière doit être portée sur chacune des étapes expérimentales impliquée, afin d'optimiser le protocole en regard de la variabilité. A titre d'exemple, HATFIELD, en 2003, liste quelques effets particulièrement sensibles, lorsque les cellules sont centrifugées puis congelées pour leur extraction ultérieure :

- ☞ une différence de température pendant la centrifugation peut déclencher le mécanisme de réponse au stress, et modifier le profil d'expression ;
- ☞ le profil d'expression peut être modifié en raison d'une petite différence de concentrations entre le tampon utilisé pour la centrifugation et le milieu de culture (réponse au stress osmotique) ;
- ☞ le métabolisme cellulaire peut être perturbé par le retrait de nutriments essentiels.

Tenant compte de ces effets indésirés, HATFIELD *ET AL.* recommandent donc d'extraire l'ARNm aussi rapidement que possible, afin de s'approcher au maximum du niveau d'expression rencontré dans les cellules et d'inhiber l'activité des ribonucléases [67].

SPRUILL *ET AL.* (2002), étudient différentes sources de variabilité, afin d'en tirer des conclusions qui permettent d'améliorer le design expérimental [126]. A cette fin, ils ont étudié l'expression de deux gènes exprimés constitutivement, codant respectivement pour la glucose-6-phosphate déshydrogénase et la bêta-actine. L'ARN utilisé provient de dix foies de cadavres. Chaque spot de la plaque (*slide*) mise au point correspond à l'ADNc de foie humain, représentés par des oligonucléotides d'une taille de 50 bases, transférés depuis une plaque 96-puits par la pipette d'un robot, sur la surface de la glace. Les sources de variabilité ont été étudiées par une procédure ANOVA, sur base du modèle présenté dans l'équation I.C.1.

$$\begin{aligned} Y_{ijklmn} = {} & L_i + S(L)_{ij} + G_k + P_l + M_m + GP_{kl} + GM_{km} + PM_{lm} + W(PM)_{lmn} \\ & + GW(PM)_{klmn} + LG_{ik} + LP_{il} + LM_{im} + LGM_{ikm} + GS(L)_{ijk} \\ & + PS(L)_{ijl} + MS(L)_{ijm} + GPS(L)_{ijkl} + E_{ijklmn} \end{aligned} \qquad \text{(Equ. I.C.1)}$$

au sein duquel les facteurs de variabilité envisagés sont les suivant:

- ☞ la plaque (Slide = S) ;
- ☞ le gène (G) ;
- ☞ la pipette du robot (P) ;
- ☞ la plaque 96-puits (Microplate = M) ;
- ☞ le foie (Liver = L) ;

- ☞ le puits (Well = W) ;
- ☞ les éventuelles interactions entre ces différents critères.

Chacun des facteurs pris en compte dans le modèle est considéré comme critère fixe [126].

Les résultats obtenus montrent que les principaux facteurs de variabilité sont le gène (6556 fois plus importante que l'erreur résiduelle), la plaque 96-puits (822 fois), le foie (590 fois), et la plaque (468 fois). Pour un même foie, la différence entre les plaques est donc de même ordre de grandeur que la différence entre les foies. De plus, selon les foies considérés, les mesures d'intensités réalisées sur les différentes plaques varient entre 5% et 1200%. Les auteurs ont découvert une mauvaise pratique expérimentale à l'origine de cette observation : le marquage des échantillons (*labeling*) a été réalisé séparément pour chaque plaque, et donc la variabilité étudiée cumule le facteur « plaque » et le facteur « marquage ». De plus, les hybridations n'ont pas été réalisées le même jour. Ces deux sources de variabilité ont contribué à l'importante variabilité observée. Enfin, l'importante variabilité consécutive à l'utilisation de plaques 96-puits différentes a été attribuée par les auteurs à un problème d'évaporation, car tous les échantillons de la première plaque 96-puits ont été utilisés avant ceux de la seconde plaque 96-puits, avec pour conséquence une durée d'évaporation plus élevée pour cette dernière [126].

L'exemple fourni par les auteurs illustre bien le problème posé par la variabilité des données issues d'expériences de *microarrays*. D'une part, une analyse adéquate doit tenir compte des différentes sources de variabilité, afin de mieux définir celle-ci. D'autre part, une telle étude fournit des arguments sur les bonnes et mauvaises pratiques liées à la technologie, à la mise au point de l'expérience, à sa mise en oeuvre par l'expérimentateur ... L'amélioration des analyses de l'expression différentielle au départ de cette technologie doit donc être envisagée par deux voies complémentaires : d'une part, en veillant à réduire la variabilité des résultats grâce à un design optimal de l'expérience et de la prise en considération des différents facteurs impliqués, et d'autre-part par l'amélioration des méthodes d'analyse statistiques et de leur capacité à estimer correctement la variabilité des données [126]. Un dialogue entre biologistes et statisticiens est donc impératif, apportant aux premiers un guide des bonnes pratiques, et aux seconds une liste des critères à prendre en compte lors de l'analyse.

I.C.3. Les méthodes de prétraitement

I.C.3.a. Introduction

Les données provenant de *microchips* se réfèrent à l'intensité lumineuse de fluorescence des *probes*. Avant d'entamer l'analyse de l'expression différentielle proprement dite, il convient de réaliser un ensemble de prétraitements statistiques sur ces données afin d'en améliorer la qualité (*low-level analyses*).

A l'instar de l'analyse de l'expression différentielle, il existe de nombreuses variétés de méthodes de prétraitement.

Les principes sur lesquels reposent ces méthodes sont parfois très différents, et elles génèrent des résultats différents, qui peuvent influencer les analyses réalisées en aval.

I.C.3.b. Prétraitements statistiques

Les méthodes présentées dans ce paragraphe proposent généralement d'effectuer un ensemble de quatre étapes de prétraitement, mais il est possible de combiner entre eux les algorithmes proposés par chaque méthode, étape par étape. Les différentes combinaisons possibles entre ces étapes et le choix des paramètres pour chacune d'entre elles, conduisent à d'innombrables possibilités de prétraitements.

Les quatre étapes principales qui peuvent être envisagées lors de ces prétraitements sont les suivantes :

- ☞ La correction du bruit de fond : L'intensité lumineuse émise par une *probe* est considérée en fonction de son hybridation mais aussi d'un bruit de fond dont l'importance varie selon la plaque, et selon la position de la *probe* sur la plaque. La correction consiste à atténuer autant que possible ce bruit de fond. Il s'agit de l'une des étapes les plus importantes du prétraitement.
- ☞ La normalisation : La normalisation vise à équilibrer les niveaux d'intensité entre les régions d'une même puce, et entre puces d'une même expérience. Il ne s'agit pas d'une normalisation au sens gaussien du terme, mais d'une standardisation du niveau d'intensité. La première étape de la normalisation corrige les données afin de

les rendre indépendante de la position des *probes* sur la puce. La seconde étape de la normalisation repose sur une standardisation des données issues de plusieurs puces différentes, de sorte que leurs distributions soient comparables.

- ☞ La correction PM/MM : La présence de *probes* MM (*MissMatch*) pour lequel un nucléotide ne permet pas un appariement correct, est utilisée pour affiner la valeur des *probes* PM (*PerfectMatch*) en tenant compte de l'hybridation non spécifique. La littérature, nombreuse et contradictoire sur ce sujet, ne permet pas de déterminer clairement les avantages et désavantages de cette correction. Certaines méthodes utilisent la correction proposée, d'autres non.

- ☞ La *summarization* : Après le prétraitement réalisé sur les données propres aux *probes*, la *summarization* combine toutes les valeurs d'intensité des *probes* relatives à un même transcrit en une seule valeur d'expression, représentative du *probeset* (gène). Bien plus qu'une simple moyenne ou addition, cette étape peut utiliser des filtres basés sur la valeur d'expression, ou des pondérations basées sur les valeurs relatives des *probes* entre elles.

I.C.3.c. Méthodes utilisées au cours de nos recherches

Le choix des méthodes de prétraitement à utiliser est important, essentiellement pour des raisons de cohérence au sein du laboratoire ou lors de collaborations extérieures, mais aussi afin de pouvoir utiliser les mêmes méthodes tout au long de ce travail.

Nous avons choisi d'utiliser les méthodes *MAS5* (ainsi que le logarithme des valeurs générées) et *GCRMA* [71, 146], sur base des données bibliographiques et des nombreux tests effectués au sein de notre unité, durant les premiers développement du projet (Berger, F. & Gaigneaux, A. - communication personnelle) ou par des partenaires (Ruelle, J.-L. - communication personnelle). Il s'agit de méthodes robustes, dont l'efficacité est avérée, selon de nombreux *benchmarks* [36, 59, 66, 116, 117].

D'autres méthodes existent et présentent de nombreuses qualités (*PLIER* [2], *DFCM* [31], *RMA*[76], *dChip*[95, 96]). Le prétraitement des données représente un sujet d'étude à part entière. Nous limiterons donc cette présentation introductive aux méthodes choisies.

La méthode RMA (*Robust Multichip Average*) n'utilise pas de correction PM/MM. La correction du bruit de fond est réalisée en utilisant un modèle basé sur la distribution

empirique des intensités de *probes*, dans lequel le signal observé est considéré comme une superposition d'un bruit de fond normal et d'un signal exponentiel. La normalisation y est effectuée en utilisant les méthodes *quantile* et *median-polish summarization*, robustes par définition (utilisation des quantiles et de la médiane).

La méthode GCRMA (GC *Robust Multi-array Average*) utilise les mêmes procédures de normalisation et de *summarization* que RMA, mais diffère par la manière de traiter le bruit de fond. GCRMA utilise la composition nucléotidique des *probes* en G et C. Cette information est utilisée pour calculer une mesure d'affinité (l'affinité entre G et C est plus forte que celle entre A et T). La distribution du bruit de fond est ensuite estimée en regroupant les *probes* dont les affinités sont semblables. Le bruit de fond est calculé spécifiquement pour chaque classe d'affinité.

Le logiciel MAS 5.0. a été développé par la société Affymetrix. La correction du bruit de fond s'y effectue de manière locale. Dans chaque région, les *probes* dont l'intensité est la plus faible (2% de la population) sont utilisées de manière pondérée pour évaluer la valeur du bruit de fond. Une correction PM/MM est effectuée et la normalisation s'effectue pour la totalité de la puce, et en tenant compte de l'ensemble des puces utilisées.

I.C.4. Au sujet de l'hypoxie...

Plusieurs jeux de données utilisés au cours de nos recherches concerne la réponse cellulaire à l'hypoxie (privation d'oxygène). Afin d'aider le lecteur, nous souhaitons mentionner brièvement les mécanismes impliqués.

Au sein des organismes qui pratiquent la respiration, l'énergie est extraite des aliments grâce à l'oxygène (O_2). Elle peut ensuite être stockée (graisses), délocalisée (glucose), ou utilisée pour créer de nouveaux composés (enzymes, ADN, composants cellulaires...), et pour diverses activités internes (digestion, immunité, contraction musculaire...). Nous ne rentrerons pas dans les détails biochimiques liés à la respiration, et à l'utilisation de l'oxygène pour extraire l'énergie des aliments. Toutefois, il nous semble important de mentionner, pour les lecteurs qui ne sont pas biologistes, que ce procédé est réalisé grâce à trois voies métaboliques fondamentales :

- ☞ la glycolyse décrit l'ensemble des réactions qui décomposent le glucose (sucre de 6 carbones) en pyruvate (3 carbones), en libérant de l'énergie. Celle-ci est utilisée pour réduire les co-facteurs NAD+ et FAD ;
- ☞ le cycle de Krebs, également appelé cycle des acides tri-carboxyliques (TCA) ou cycle du citrate : convertit le pyruvate successivement sous forme de plusieurs acides tricarbonés, et utilise le produit final pour reformer du pyruvate. Ce cycle est une plaque tournante du métabolisme : les produits de dégradation des protéines et des graisses sont des composés intermédiaires du cycle de Krebs, inscrits également dans plusieurs voies de biosynthèses ;
- ☞ la phosphorylation oxydative désigne l'ensemble des réactions qui permettent d'utiliser l'oxygène pour récupérer l'énergie stockée temporairement dans les cofacteurs et la convertir en une forme utilisable. Au cours de cette opération, les cofacteurs (NADH et FADH2) sont oxydés, ce qui crée une différence de concentration en H+ entre la mitochondrie et le cytoplasme. Ceci-ci est utilisée par des protéines membranaires pour former de l'ATP au départ d'ADP et de phosphate. L'ATP est donc la molécule utilisée pour stocker l'énergie, grâce à l'oxygène respiré.

Ces trois voies métaboliques sont interconnectées avec le métabolisme des protéines et des

lipides. De plus, l'activité cellulaire implique des voies de signalisation, capables de détecter le manque d'oxygène, et de réagir en régulant l'activité cellulaire.

II. Analyse de l'expression différentielle

II.A.
Analyse par gène

Résumé

Ce chapitre présente un échantillon non exhaustif des méthodes actuelles d'analyse de l'expression individuelle. Les méthodes sélectionnées fournissent toutefois un aperçu représentatif des différentes stratégies publiées.

L'examen des différentes procédures disponibles au terme de ce travail montre que plusieurs études convergent, et formulent mathématiquement l'idée de partager de l'information entre les gènes. Au départ des méthodes classiques (*t* de STUDENT, correction de WELCH), ils décrivent des modèles bayesiens qui reposent sur l'estimation d'un ou plusieurs paramètres *a priori*, calculé sur l'ensemble ou une catégorie de gènes. Les méthodes bayesiennes d'estimation de la variance, ainsi que l'utilisation d'un estimateur de STEIN, conduisent à une formulation similaire de la correction de la variance, ainsi que nous le montrerons également via une formulation généralisée dans la première partie du chapitre Résultats. Les deux principales distinctions entre ces méthodes sont la manière dont la correction est pondérée, et le critère choisi pour déterminer le(s) paramètre(s) de correction utilisé(s) *a priori*.

Les différentes méthodes présentées dans ce chapitre peuvent être classées en trois catégories : les méthodes dites « classiques », paramétriques ou non paramétriques, et les méthodes bayesiennes ou quasi-bayesiennes. Parmi ces dernières, les critères utilisés comme *a priori* sont aussi variés que les méthodes et modèles définis.

Ainsi, chronologiquement, les méthodes d'analyse individuelles ont évolué, depuis l'utilisation de la simple comparaison de deux valeurs, en passant par la comparaison de deux séries de valeurs, pour finalement s'adapter en utilisant plusieurs approches pour améliorer les performances de la comparaison de séries de valeurs, sur base du partage d'information entre les gènes.

II.A.1. Méthodes classiques

II.A.1.a. Le k-fold

La méthode la plus simple pour déterminer les gènes significativement sur- ou sous-exprimés repose sur le rapport entre les données d'expression relatives au témoin et celles relatives aux conditions testées. Si ce rapport, également appelé *fold change*, est supérieur à une valeur seuil choisie (k), le gène est considéré comme différentiellement exprimé (Equation II.A.1) [43].

$$r = \frac{\mu_2}{\mu_1} \quad \text{(Equ. II.A.1)}$$

avec μ_1, μ_2, les moyennes des valeurs dans les deux échantillons, et r, le rapport qui sera comparé au seuil k.

Cette méthode historique présente plusieurs inconvénients. Sa principale limitation réside dans la difficulté de choisir une valeur de seuil k adéquate. En effet, ce rapport dépend non seulement du niveau d'expression envisagé, mais aussi de la variabilité des données propres à chaque gène. Ainsi, une valeur seuil donnée peut fournir de bons résultats pour un niveau d'expression donné, mais peut aussi, appliquée à un autre niveau d'expression, conduire à des erreurs de type I (faux positifs, FP) et II (faux négatifs, FN) assez importantes. De même, le niveau d'expression d'un gène donné est variable. En raison de cette variabilité, il est nécessaire de définir un seuil k élevé pour éviter de sélectionner des gènes dont la valeur r est élevée par hasard. De plus, une faible différence du niveau d'expression, associée à une faible variabilité, risque de ne pas être détectée (erreur de type II). A l'inverse, un seuil plus petit permet de détecter ces gènes, mais ils seront dilués parmi les gènes de niveau d'expression plus variable mais non régulés (erreur de type I).

II.A.1.b. Test de STUDENT

La mesure de la différence entre les moyennes des échantillons est étudiée en relation avec la déviation standard associée grâce au test statistique de STUDENT [130]. Pour chaque gène, la statistique t est définie sur base de l'équation II.A.2.

$$t_i = \frac{\mu_{i,1} - \mu_{i,2}}{\sqrt{\frac{\sigma_{i,1}^2}{n_1} + \frac{\sigma_{i,2}^2}{n_2}}} \quad \text{(Equ. II.A.2)}$$

$$df = n_1 + n_2 - 2$$

avec $\mu_{i,1}$, $\sigma_{i,1}^2$, et n_1, la moyenne, la variance et la taille du premier échantillon, et $\mu_{i,2}$, $\sigma_{i,2}^2$ et n_2, la moyenne, la variance et la taille du second échantillon. t_i est la statistique de STUDENT associée au gène i, et df symbolise les degrés de libertés associés au test.

Le test de STUDENT s'applique lorsque les données suivent une distribution normale, et présentent une variance homogène entre les séries d'observations indépendantes [130].

Cette méthode est appropriée pour détecter les gènes différentiellement exprimés. En effet, comme le montre l'équation II.A.2, elle privilégie les gènes pour lesquels la différence entre les niveaux d'expression entre les deux groupes est maximale et la variation de l'expression au sein de chaque groupe est minimale.

Le nombre de réplicats relatifs à une même condition expérimentale est souvent réduit, ce qui aboutit à une estimation peu fiable de la variance et donc à une diminution des performances.

GOLUB *ET AL.* (1999) proposent une alternative similaire au test du *t* de STUDENT basée sur le rapport signal/bruit (SNR ou S2N) [65]. La mesure de corrélation entre les échantillons y est définie par l'équation II.A.3.

$$SNR_i = \frac{m_{i,1} - m_{i,2}}{s_{i,1} - s_{i,2}} \quad \text{(Equ. II.A.3)}$$

où $m_{i,1}$ et $s_{i,1}$ sont la moyenne et la déviation standard du gène i dans la première

condition, $m_{i,2}$ et $s_{i,2}$ sont la moyenne et la déviation standard du gène i dans la seconde condition.

II.A.1.c. Correction de Welch pour le test du t de Student

Une variante du test du *t* de STUDENT, proposée par WELCH permet de comparer entre eux des groupes dont la variance est hétérogène [143]. Le calcul des degrés de liberté y est adapté, de façon similaire à l'approximation de SATTERTHWAITE (Equations II.A.4 et II.A.5) [121, 143].

$$t_i = \frac{\mu_{i,1} - \mu_{i,2}}{\sqrt{\frac{\sigma_{i,1}^2}{n_1} + \frac{\sigma_{i,2}^2}{n_2}}} \quad \text{(Equ. II.A.4)}$$

$$df = \frac{\frac{\sigma_{i,1}^2}{n_1} + \frac{\sigma_{i,2}^2}{n_2}}{\frac{\left(\frac{\sigma_{i,1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\sigma_{i,2}^2}{n_2}\right)^2}{n_2 - 1}} \quad \text{(Equ. II.A.5)}$$

où $\sigma_{i,1}^2$ et n_1 sont la variance et la taille du premier échantillon, $\sigma_{i,2}^2$ et n_2 sont la moyenne, la variance et la taille du second échantillon. t_i est la statistique de STUDENT associée au gène i, et df symbolise les degrés de libertés associés au test.

Les conditions d'application sont identiques à celles du test du *t* de STUDENT (distribution normale et observations indépendantes). Ce test est défini pour deux échantillons de variances hétérogènes [143]. Il apparaît donc approprié à l'analyse de l'expression différentielle au départ des données issues de *microchips*.

II.A.1.d. Test de la somme des rangs

Le test de Student est qualifié de « paramétrique », car il repose sur l'usage d'indicateurs intermédiaires, la moyenne et la variance, qui sont ensuite comparés à une distribution de référence (la distribution normale). Le test de STUDENT, en revanche, est inutilisable si les données observées ne suivent pas une distribution normale. D'autres méthodes ont été développées pour permettre l'analyse de telles données.

Une stratégie régulièrement utilisée consiste à attribuer un score à chaque mesure réalisée, et utiliser ensuite ce score par comparaison avec une attribution aléatoire de scores. L'équivalent non paramétrique du test de STUDENT est le test de la somme des rangs, décrit par WILCOXON et MANN & WHITNEY dans deux études distinctes [103, 145].

La première étape consiste à attribuer un rang à chaque mesure des deux conditions comparées. La somme des rangs attribués est ensuite évaluée séparément pour les deux conditions. Si les deux séries de mesures comparées appartiennent à la même population, les deux valeurs obtenues sont proches les unes des autres. Des valeurs forts différentes peuvent être observées par hasard, avec une fréquence qui peut être évaluée par comparaison à une distribution aléatoire de scores. Si les deux échantillons comparés sont différents, alors des scores plus petits sont attribués à l'une des séries de mesures, et des scores plus grands dans l'autre. Il suffit de comparer la valeur de la somme d'une des deux séries de mesures avec ses valeurs possibles pour lui attribuer une *p-value*.

La significativité du test est donc réalisée sur base de permutations des scores. Lorsque le nombre de mesures disponible est élevé, il n'est cependant pas nécessaire d'évaluer toutes les permutations possibles, opération très exigeante en terme de puissance de calcul, car les résultats tendent vers une distribution normale lorsque le nombre de mesures augmentent. Celle-ci peut donc être utilisée, comme dans les tests paramétriques, pour évaluer la significativité du test. Les relations mathématiques fournies dans les équations II.A.6 et II.A.7 sont utilisées pour caractériser la distribution normale utilisée comme référence.

$$\mu = \frac{n_1\left(n_1+n_2-1\right)}{2} \quad \text{(Equ. II.A.6)}$$

$$\sigma = \sqrt{\frac{n_1 n_2\left(n_1+n_2+1\right)}{12}} \quad \text{(Equ. II.A.7)}$$

ou μ et σ sont les moyennes et variances de la distribution de la somme des rangs. n_1 et n_2 désignent la taille des deux échantillons comparés.

Lorsque le nombre de mesures disponibles est limité, l'usage de la distribution normale doit être remplacé par des permutations, ce qui implique une discrétisation de la distribution des *p-values* évaluées, et l'usage du test de WILCOXON et MANN & WHITNEY implique l'attribution de la même *p-value* à de nombreux gènes [103, 145].

II.A.1.e. Le test du produit des rangs

BREITLING *ET AL.*, en 2004, ont décrit une procédure non paramétrique mise au point pour analyser les données d'expression. Leur méthode, dénommée « *Rank Products* », partage l'utilisation de scores avec le test de WILCOXON. Ces scores sont cependant attribués et évalués différemment [24].

Dans une première étape, les auteurs utilisent le rapport des données d'expression comparées, pour chaque comparaison deux à deux possible. Chaque gène se voit attribuer une liste de *fold changes* (FC), rassemblée dans une nouvelle matrice. Chaque colonne de la matrice se réfère à une comparaison deux à deux, et chaque ligne se réfère à un gène.

Pour chaque comparaison évaluée, la procédure consiste ensuite à attribuer un score à chaque gène, en fonction de la valeur de FC qui lui est associé pour cette comparaison. Une nouvelle matrice de scores est ainsi obtenue. La procédure est donc compétitive, car elle évalue chaque gène par comparaison avec les autres gènes.

Les valeurs disponibles pour chaque gène dans la matrice des scores sont ensuite utilisées individuellement. Le produit des rangs associés à chaque gène est calculé, et une valeur élevée suggère une expression différentielle. La significativité du test est évaluée sur base de permutations des scores, pour chaque comparaison deux à deux. La distribution du produit des rangs obtenue par permutations est utilisée comme distribution de référence, pour compter le nombre de réalisations dues au hasard [24].

II.A.2. Méthodes bayesiennes et quasi-bayesiennes

II.A.2.a. Introduction

L'approche statistique bayesienne est probabiliste: les probabilités y sont interprétées comme un degré de confiance en une hypothèse plutôt qu'en fréquences mesurées (par opposition à l'approche fréquentiste). Le modèle de BAYES permet de mettre à jour ces probabilités lors de l'apport de nouvelles données [14, 15].

Le théorème bayesien évalue la probabilité *a posteriori* P d' une hypothèse H, étant donné les données D de la manière suivante :

$$P(H|D)=\frac{P(D|H)P(H)}{P(D)} \quad \text{(Equ. II.A.8)}$$

Où $P(H)$ est la probabilité *a priori* d'acceptation de l'hypothèse nulle, sans considérer les informations utilisables en complément [14, 15].

La méthode du *shrinkage t,* décrite au paragraphe II.A.2.k. , est qualifiée de quasi-bayesienne par ses auteurs. En réalité, elle ne repose pas sur un modèle bayesien mais conduit à une formulation similaire de la statistique individuelle étudiée, sur base d'un estimation dérivant du modèle de JAMES STEIN [109]. La méthodologie LPE est également décrite dans cette section (paragraphe II.A.2.c.), car elle partage avec les méthodes bayesiennes l'utilisation de données empiriques supplémentaires pour guider l'analyse statistique [77].

II.A.2.b. Le « regularized t-test »

L'un des problèmes rencontrés lors de l'analyse de l'expression différentielle sur base du test de STUDENT est la définition d'un estimateur approprié de la variance. HUNG, en 2002, a montré qu'une relation existe entre la variance et le niveau d'expression génique. Cette connaissance préliminaire peut être utilisée en conjonction avec une approche bayesienne pour estimer la variance individuelle en examinant le niveau d'expression des gènes [75].

BALDI *ET AL.* ont décrit un modèle bayesien, reposant sur une distribution normale caractérisée par sa moyenne μ et sa variance σ. Les auteurs utilisent une distribution de

Dirichlet pour le paramètre conjugué *a priori*, et σ^2 suit une distribution gamma inverse (*scaled inverse gamma*). Le modèle proposé est caractérisé par les équations II.A.9 et II.A.10 [11, 67, 99].

$$P(\mu|\sigma^2) = N\left(\mu ; \mu_0, \frac{\sigma^2}{\lambda_0}\right) \text{ (Equ. II.A.9)}$$

$$P(\sigma^2) = I\left(\sigma^2 ; \nu_0, \sigma_0^2\right) \text{ (Equ. II.A.10)}$$

Le modèle se base sur la définition d'un vecteur comportant 4 hyper-paramètres, $\alpha = (\mu_0, \lambda_0, \nu_0, \sigma_0^2)$. μ_0 et σ^2/λ_0 sont interprétés comme des paramètres de localisation et d'échelle de μ, et ν_0 et σ_0^2 sont les degrés de liberté l'échelle de σ^2.

L'application du théorème de BAYES conduit à l'équation II.A.11, similaire aux équations II.A.9 et II.A.10.

$$P(\mu, \sigma^2|D, \alpha) = N\left(\mu ; \mu_n, \frac{\sigma^2}{\lambda_n}\right) I\left(\sigma^2 ; \nu_n, \sigma_n^2\right) \text{ (Equ. II.A.11)}$$

avec

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n}\mu_0 + \frac{n}{\lambda_0 + n}m$$

$$\lambda_n = \lambda_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n \sigma_n^2 = + \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\lambda_0 n}{\lambda_0 + n}(m - \mu_0)^2$$

Les paramètres *a posteriori* sont dérivés des paramètres *a priori* et des données traitées. Ainsi, dans l'équation II.A.11,

- ☞ μ_n est une moyenne pondérée du paramètre initial μ_0 et de la moyenne de l'échantillon ;
- ☞ ν_n est défini par les degrés de libertés *a priori* ν_0, auquel et par la taille de

l'échantillon (n);

☞ $\nu_n\sigma_n^2$, la somme des carrés des écarts *a posteriori*, est la somme des carrés des écarts *a priori* $(\nu_0\sigma_0^2)$, à laquelle s'ajoute la somme des carrés des écarts de l'échantillon $(n-1)s^2$ et une incertitude résiduelle due à la différence entre la moyenne μ_0 et la moyenne de l'échantillon m, formulée par $(\lambda_0 n/(\lambda_0+n))(m-\mu_0)^2$.

L'estimation des paramètres du modèle bayesien, sur base de la moyenne *a posteriori* (MP) ou du maximum *a posteriori* (MAP), aboutissant aux équations II.A.12 et II.A.13.

$$\mu=\mu_n \text{ et } \sigma^2=\frac{\nu_n}{\nu_n-2}\sigma_n^2 \text{ (Equ. II.A.12)}$$

$$\mu=\mu_n \text{ et } \sigma^2=\frac{\nu_n}{\nu_n-1}\sigma_n^2 \text{ (Equ. II.A.13)}$$

En posant $\mu_0=m$, nous obtenons les équations II.A.14 et II.A.15.

$$\mu=m \text{ et } \sigma^2=\frac{\nu_n}{\nu_n-2}\sigma_n^2=\frac{\nu_0\sigma_0^2+(n-1)s^2}{\nu_0+n-2} \text{ (Equ. II.A.14)}$$

$$\mu=m \text{ et } \sigma^2=\frac{\nu_n}{\nu_n-1}\sigma_n^2=\frac{\nu_0\sigma_0^2+(n-1)s^2}{\nu_0+n-1} \text{ (Equ. II.A.15)}$$

La variance empirique y apparaît modulée par ν_0 observations supplémentaires et par une variance σ_0^2, estimée sur l'ensemble des données ou sur une partie de celles-ci. Par défaut, sa valeur est calculée sur une fenêtre de 50 gènes situés de part et d'autre du gène d'intérêt dans un tri basé sur le niveau d'expression (101 *probesets* au total)[11]. s^2 est l'estimateur individuel de la variance.

L'estimation de ν_0 est délicate. HATFIELD propose d'attribuer à ν_0+n une valeur constante, par exemple égale à 10 [67]. Cette valeur est celle qui est utilisée par défaut dans le logiciel *CyberT*, reposant sur le *regularized t-test* [11].

Les estimateurs de la tendance centrale et de la dispersion obtenus sont ensuite introduits dans le test de STUDENT, en incluant la correction de WELCH. Les conditions d'application

sont celles du test de STUDENT et de l'existence d'une relation entre le niveau d'expression et la variabilité.

L'addition d'information utilisée pour estimer la variance permet à certains auteurs (HATFIELD, en 2003 [67]), d'affirmer que cette approche bayesienne fourni une estimation adaptée de la variance, et permet de réduire le *false discovery rate* (FDR) pour les expériences caractérisée par un nombre réduit de mesures [11, 99].

Sur base de cette méthodologie, FOX, en 2006, soulève plusieurs critiques relatives au nombre de degrés de libertés, ν_0+n, utilisés par BALDI. D'une part, étant donné que les variances sont estimées séparément pour les deux échantillons, le nombre total de degrés de libertés utilisés par le *regularized t-test* vaut $n_1+n_2+2\nu_0-4$, conduisant à n_1+n_2-4 lorsque $\nu_0=0$ (la fenêtre n'est pas utilisée), ce qui n'est pas en accord avec le test du t réalisé classiquement, caractérisé par n_1+n_2-2. D'autre part, dans la stratégie suivie par BALDI & LONG, ν_0 est fixé arbitrairement à 10, et la fenêtre utilisée pour estimer s_0 comprends 101 *probesets*. FOX critique l'indépendance de ces deux paramètres, et ajuste la stratégie en imposant un lien entre ν_0 et la taille de la fenêtre utilisée, en suivant l'équation II.A.16. En conséquence, il préconise d'utiliser une fenêtre dont la taille est fixée par le paramètre ν_0, en accord avec la formulation mathématique de cet hyper-paramètre. Le nombre de gènes considérés dans la stratégie ajustée de FOX est par conséquent beaucoup plus petit, et en relation avec le nombre de réplicats. A titre d'exemple, pour $n=2$ et $\nu_0=8$, 8 gènes sont utilisés pour estimer la variance *a priori*, au lieu de 101 dans la stratégie suivie par BALDI & LONG. De plus, FOX exclus le gène d'intérêt de la fenêtre, et ne sélectionne que les gènes voisins pour un même niveau d'expression [58].

$$\nu_0=m(n-1) \quad \text{(Equ. II.A.16)}$$

avec m, le nombre de gènes inclus dans la fenêtre, et n, le nombre de mesures réalisées.

II.A.2.c. Le test LPE (Local Pooled Error)

La méthode *LPE*, décrite par N. JAIN (2003), repose également sur l'utilisation de la relation empirique entre le niveau d'expression et la variabilité. Bien qu'il ne s'agisse pas d'un modèle bayesien, le principe de la méthode est similaire au *regularized t-test*, car les auteurs utilisent la relation empirique observée comme *a priori* pour prédire la variabilité

individuelle [77].

Pour modéliser la relation entre le niveau d'expression et la variabilité, une procédure en deux étapes est utilisée. La première étape consiste à définir des intervalles réguliers sur base de la médiane (définis par les quantiles de la distribution). La seconde étape affine la démarche en utilisant des intervalles de tailles variables, en fonction de la variabilité propre à chaque niveau d'expression [77].

Les données d'expression utilisées par le test *LPE* sont tout d'abord utilisées pour évaluer les statistiques A et M, par comparaison deux à deux de toutes les données d'expression d'un même *probeset* (gène) entre les deux conditions comparées, sur base des équations II.A.17 et II.A.18. La statistique A est représentative du niveau d'expression, et la statistique M est représentative des écarts observés entre les mesures [77].

$$A = \frac{X_1 + X_2}{2} \quad \text{(Equ. II.A.17)}$$

$$M = \left| X_1 - X_2, X_2 - X_1 \right| \quad \text{(Equ. II.A.18)}$$

où X_1 et X_2 dénotent les valeurs individuelles associées à un jeu de données comportant par exemple deux réplicats. Lorsque le jeu de données comporte davantage de mesures répétées, les statistiques A et M sont calculées pour chaque comparaison deux à deux possible. Les statistiques A et M sont ensuite utilisées, pour chaque condition testée, pour estimer la variance individuelle par interpolation [77].

Au cours de la première étape, la statistique A permet de déterminer 100 intervalles définis par les percentiles de la distribution. La relation de référence entre la moyenne et la variance est ensuite caractérisée en calculant la médiane de A dans chaque intervalle et la variance de M associée à un terme de correction de la variance F_m, calculé sur base de l'équation II.A.19 :

$$F_m = \frac{0.5\left(n_k - 0.5\right)}{n_k - 1} \quad \text{(Equ. II.A.19)}$$

où k est l'indice de l'intervalle considéré et n_k est le nombre de valeurs contenues dans l'intervalle k.

Les valeurs de la médiane et de la variance associées à chaque intervalle fournissent les coordonnées (A_{med}, σ^2) de la relation entre le niveau d'expression et la variabilité. La

courbe obtenue fait ensuite l'objet d'un lissage. Sur base de la courbe obtenue, il est possible d'interpoler la variance individuelle associée à chaque valeur A (dénotée s^2) [77].

La seconde étape est similaire. La différence essentielle entre les deux étapes repose sur le nombre de valeurs utilisées pour calculer les coordonnées associées à chaque intervalle. Des intervalles de tailles variables sont définis, contenant chacun plus de dix valeurs, et moins de 1% de l'ensemble des valeurs, sur base de la variabilité associée à chaque intervalle défini. Le premier intervalle est défini par toutes les valeurs comprises entre la valeur de la statistique A associée au premier gène de la liste triée et $A_1+\sqrt{s^2}$. Les intervalles suivants sont définis de la même manière à partir de la première valeur de A située en dehors de l'intervalle. Les dernières valeurs de la liste sont inclues dans le dernier intervalle. Les intervalles définis sont alors utilisés, à l'instar de la première étape, pour calculer les coordonnées (A_{med}, σ^2) et estimer la variance des gènes par interpolation [77].

L'analyse de l'expression différentielle par cette méthode repose ensuite sur la définition de la statistique Z. Celle-ci est définie classiquement par le rapport entre d'une part la différence des valeurs médianes associées à chaque condition, et d'autre part la variance obtenue par interpolation sur le graphe obtenu à l'issue de la seconde étape. La significativité du résultat est ensuite évaluée sur base d'une distribution normale [77].

Bien que l'idée de l'utilisation de la relation empirique entre le niveau d'expression et la variabilité soit partagée avec le *regularized t-test*, plusieurs éléments essentiels distinguent ces deux procédures :

- ☞ La relation est utilisée sur base de la moyenne dans le *regularized t-test*, et sur base de la médiane dans le test *LPE* ;
- ☞ La relation de référence est utilisée implicitement dans le *regularized t-test*, en utilisant les gènes voisins. Elle est utilisée explicitement dans le test *LPE* en utilisant un lissage de la courbe pour interpoler des valeurs ;
- ☞ La valeur finale de la variance associée à chaque niveau d'expression, dans le test *LPE*, est interpolée sur base d'intervalles définis avec une taille variable, d'autant plus grand que les données sont similaires. La taille de la fenêtre utilisée par le *regularized t-test* est constante ;
- ☞ La significativité est évaluée sur base d'une distribution normale en lieu et place

d'une distribution *t*. La statistique utilisée s'intéresse à la différence des valeurs médianes au lieu de la différence des valeurs moyennes ;

☞ Aucun modèle mathématique n'est utilisé pour pondérer les informations individuelles et la relation avec le niveau d'expression dans le test *LPE*. La variance individuelle est simplement « prédite » sur base de la courbe de référence obtenue empiriquement.

II.A.2.d. Modèle bayesien hiérarchique et catégorisation de la variance

Plusieurs des méthodologies évoquées dans cette introduction considèrent des modèles ou deux populations de gènes sont représentés, les gènes non impliqués, et les gènes différentiellement exprimés. Dans une approche similaire, DELMAR *ET AL.* utilisent un modèle de définition de populations différentes sur base de l'homogénéité de la variance. Chaque gène se voit ensuite attribuer une variance en relation avec la classe à laquelle il appartient, estimée sur un grand nombre de gènes [39].

MANDA *ET AL. (*2007), partant d'une idée similaire, proposent d'utiliser un modèle bayesien hiérarchique, permettant de modéliser les sources de variabilité dans un modèle commun, et de propager l'information. Dans cette approche, les variances sont stabilisées et pour chaque catégorie de variance définie, la variance individuelle est utilisée en conjonction avec la variance liée à la catégorie d'appartenance. La procédure d'analyse a été implémentée dans le logiciel *WinBUGS* [102].

Le modèle présenté repose sur un jeu de données où deux expériences sont co-hybridées avec un échantillon de référence. En cas d'acceptation de l'hypothèse nulle, les moyennes des échantillons comparés entre les deux expériences sont identiques, et la variance est une variable χ^2 avec ν degrés de liberté (équation II.A.20), dont la densité est définie par l'équation II.A.21 [102].

$$w_i = \nu S_i^2 \sim \Psi_i^{-1} \chi_\nu^2 \quad \text{(Equ. II.A.20)}$$

$$g(w_i|\Psi_i, \nu) = \frac{(\Psi_i)^{\nu/2}}{2^{\nu/2}\Gamma(\nu/2)} w_i^{\nu/2-1} \exp(-\frac{1}{2}\Psi_i w_i) \quad \text{(Equ. II.A.21)}$$

La moyenne de w_i vaut ν/Ψ_i, et la moyenne des variances individuelles, S_i^2, vaut $1/\Psi_i$. La constante Ψ_i est inconnue et doit être déterminée au départ des données. Plutôt que de

l'estimer pour chaque gène, MANDA *ET AL.* préconisent de modéliser w_i au départ d'observations indépendantes d'un mélange de plusieurs distributions χ^2 (Equation II.A.22) [102].

$$g(w_i|\pi,\Psi,v)\sim\sum_{j=1}^{k}\pi_j g(w_i|\Psi_j,v) \quad \text{(Equ. II.A.22)}$$

où k est le nombre de classes définies pour la distribution de la variance, et π_j est la proportion de gènes appartenant à la classe j.

Pour compléter le modèle, les auteurs définissent une variable d'allocation Z, réalisation indépendante de variables aléatoires discrètes dont la probabilité est définie par l'équation II.A.23. En conséquence, les valeurs observées de w_i sont redéfinies comme étant des observations indépendantes de densité définie par l'équation II.A.24, et le modèle de mélange de plusieurs distributions χ^2 s'exprime par l'équation II.A.25, identique à l'équation II.A.22 [102].

$$P(Z_i=j|\pi,\Psi,v)=\pi_i(i=1,\ldots,g\,;\,j=1,\ldots,k) \quad \text{(Equ. II.A.23)}$$

$$g(w_i|Z_i=j,\pi,\Psi,v)\sim g(w_i|\Psi_j,v) \quad \text{(Equ. II.A.24)}$$

$$g(w_i|\pi,\Psi,v)=\sum_{j=1}^{k}P(Z_i=j|\pi,\Psi,v)g(w_i|Z_i=j,\pi,\Psi,v)=\sum_{j=1}^{k}\pi_j g(w_i|\Psi_j,v) \quad \text{(Equ. II.A.25).}$$

Sur base de ce modèle, l'appartenance des gènes aux différentes classes de variances, caractérisées par π_{ij}, est déterminée par la valeur la plus élevée de la probabilité associée à la distribution de la variable d'allocation Z, calculée par l'équation II.A.26. Tous les gènes appartenant à la classe j partage la même variance de valeur $1/\Psi_j$ [102].

$$\pi_{ij}=P(Z_i=j|w_i,\pi,\Psi,v)=\frac{\pi_j g(w_i|\Psi_j,v)}{\sum_{l=1}^{k}\pi_{il}(w_i|\Psi_l,v)} \quad \text{(Equ. II.A.26)}$$

A l'inverse de DELMAR *ET AL.*, l'estimation *a priori* des paramètres π et Ψ repose sur une structure bayesienne hiérarchique, qui fait intervenir une distribution gamma et une distribution de Dirichlet, respectivement, pour estimer Ψ_j et π [39, 102].

II.A.2.e. La méthode EBAM (Empirical Bayes Analysis of Microarray data)

EFRON et ses collaborateurs ont décrit une méthode bayesienne de sélection des gènes différentiellement exprimés reposant sur la comparaison de la distribution des différences d'expression avec une distribution nulle établie empiriquement [49].

La procédure utilisée repose sur la statistique Z, définie par l'équation II.A.27. La distribution nulle de la statistique Z est établie sur base de permutations.

$$Z_i = \frac{\overline{D_i}}{a_0 + S_i} \quad \text{(Equ. II.A.27)}$$

La distribution nulle des valeurs Z, notée $f_0(z)$, est utilisée en conjonction avec la distribution des valeurs Z obtenue pour les gènes affectés, notée $f_1(z)$. Celle-ci est évaluée sur base de la comparaison des valeurs Z observées (distribution $f(z)$) avec les valeurs attendues sur base de la distribution nulle, selon l'équation II.A.28.

$$f(z) = p_0 f_0(z) + p_1 f_1(z) \quad \text{(Equ. II.A.28)}$$

Dans cette équation, p_0 représente la probabilité d'acceptation de l'hypothèse nulle pour un gène, et p_1 la probabilité d'expression différentielle. Selon la règle de BAYES, nous pouvons extraire de l'équation II.A.28 les valeurs de ces probabilités p_0 et p_1 pour un gène donné (Equation II.A.29).

$$p_1(Z) = 1 - \frac{p_0 f_0(Z)}{f(Z)} \quad \text{et} \quad p_0(Z) = \frac{p_0 f_0(Z)}{f(Z)} \quad \text{(Equ. II.A.29)}$$

Le rapport $f_0(Z)/f(Z)$ est évalué sur base des données disponibles, au lieu d'une distribution normale ou d'une distribution gamma.

L'équation II.A.27 implique un paramètre correctif a_0, qui est choisi de façon à optimaliser la séparation entre les distributions $f(Z)$ et $f_0(Z)$. La valeur utilisée par les auteurs correspond au 90ème percentile des écarts types observés [49].

Variances homogènes et observations indépendantes sont les pré-requis habituels nécessités par cette méthode.

L'avantage majeur de la méthode EBAM est de contrôler le FDR (*False Discovery Rate*)

plutôt que le FWER (*Family Wise Error Rate*). Il semble aussi moins sévère que les autres méthodes de corrections, proposées par BONFERRONI, WESTFALL & YOUNG ou BENJAMINI & HOCHBERG. A l'instar du *regularized t-test*, cette méthode élimine un certain nombre de faux positifs caractérisés par de faibles variances individuelles [49].

EFRON & TIBSHIRANI (EFRON *ET AL.*, 2002) proposent une version d'EBAM modifiée qui utilise la statistique de la somme des rangs de WILCOXON à la place de la statistique de *t* modifiée [51].

II.A.2.f. La méthode SAM

La méthode SAM, décrite en 2001 par TUSHER *ET AL.*, est dérivée de la méthode proposée par EFRON *ET AL.*[49]. La statistique étudiée repose également sur expression mathématique similaire au *t* de STUDENT. La statistique dérivée, formulée par l'équation II.A.30, est symbolisée par d [136].

$$d_i = \frac{\mu_1 - \mu_2}{s_0 + s_i} \quad \text{(Equ. II.A.30)}$$

$$\text{avec} \quad s_i = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2}\left(SSE_1 + SSE_2\right)$$

L'un des éléments les plus importants de la méthode consiste à déterminer la valeur du paramètre s_0, également appelé *fudge factor*. Sa définition permet d'éviter qu'une faible variance conduise à une statistique d très élevée, impliquant de nombreux faux positifs. La recherche de la valeur adéquate de ce paramètre repose sur le calcul de la variabilité individuelle au sein du jeu de données. Les percentiles de la distribution des valeurs individuelles (s_i) sont notés s_α, et définissent 100 intervalles de valeurs. L'étape suivante consiste à estimer la statistique d_i pour chaque gène, en remplaçant s_0 par chaque s_α défini. La valeur choisie finalement pour s_0 correspond à la valeur s_α pour laquelle le coefficient de variation de d est minimal [136].

A l'instar de la méthode EBAM, la distribution nulle des valeurs d_i est évaluée sur base de permutations. Pour chaque gène, la statistique d est calculée sur base de permutations

aléatoire des échantillons (valeurs d_i^p). La moyenne des valeurs d_i^p associées aux permutations fourni une valeur attendue, dénotée d_i^E. Les valeurs d_i calculées pour chaque gène sont ensuite triées de la plus élevée à la plus faible, de même que les valeurs d_i^E. Un graphique de d_i en fonction de d_i^E montre que la distribution des valeurs de d observées suit linéairement la distribution des valeurs de d attendues, et s'en écarte aux valeurs extrêmes. La définition du seuil de sélection s'illustre en représentant deux droites parallèles à la diagonale, et les gènes pour lesquels les valeurs d_i se situent au-delà de l'intervalle défini par les deux droites sont sélectionnés [136].

L'homogénéité de la variance et l'indépendance des observations constituent les conditions de validité du test.

Les avantages de la méthode implémentée dans SAM sont identiques à ceux de la méthode EBAM. La méthode peut aisément être étendue à de nombreuses stratégies expérimentales impliquant les *microarrays* [136].

Le nombre de mesures réalisées détermine le nombre de permutations disponibles pour estimer correctement la significativité et le FDR.

II.A.2.g. La méthode SAM améliorée : modèle de régression linéaire pénalisée

La méthode SAM repose sur la statistique d, définie par la formule II.A.30. Sur base des travaux de TIBSHIRANI et de EFRON [51, 52], WU applique un modèle de régression linéaire pénalisée, et définit ainsi la statistique t^*, une version pénalisée du *t* de STUDENT (équation II.A.31). L'expression de t^* prend la même forme que la statistique d utilisée dans la méthode SAM [147].

$$t_i^* = sign\left(\bar{x}_{i1} - \bar{x}_{i2}\right) \frac{\left(\left|\bar{x}_{i1} - \bar{x}_{i2}\right| - \lambda\right)_+}{\sqrt{\frac{n}{n_1 n_2} s_i^2 + \frac{1}{n-2}\lambda^2}} \quad \text{(Equ. II.A.31)}$$

L'ajout de la constante $\lambda^2/(n-2)$ au dénominateur joue le même rôle que s_0 dans la méthode SAM : stabiliser la variance pour tous les gènes. En revanche, dans la statistique pénalisée t^*, la différence des moyennes est diminuée d'une constante λ. Lorsque s_i et s_0 sont proches de 0, cette pénalité réduit la différence des moyennes et évite par

conséquent l'obtention de valeur élevées de t^*.

La recherche du paramètre λ est similaire à la procédure décrite par TUSHER (SAM, [136]). La comparaison des relations II.A.30 et II.A.31 conduit à l'équation II.A.32 :

$$\lambda = s_0\sqrt{n-2} \quad \text{(Equ. II.A.32)}$$

Les performances de cette méthode sont comparables à celle de SAM lorsque λ est estimé suivant l'équation II.A.32. Cependant, le paramètre λ peut être calculé sur base du modèle linéaire présenté dans l'équation II.A.33.

$$t_i^* = \beta_0 + \beta_1 s_i + \varepsilon_i \quad \text{(Equ. II.A.33)}$$

caractérisé par un coefficient (β_1) représentatif de la contribution de s_i sur la variabilité de t_i^*. Si $\beta_1 = 0$, t_i^* et s_i sont indépendants.

La valeur de λ est donc recherchée en minimisant la valeur absolue de R, le coefficient de corrélation entre t_i^* et s_i [147].

Enfin, une troisième approche est proposée pour estimer la valeur de λ, sur base d'un modèle local de régression linéaire. En considérant une fenêtre autour d'un gène, t_i^* et s_i sont ajustés sur un modèle linéaire ou polynomial. L'erreur résiduelle reflète dès lors la dépendance entre t_i^* et s_i. La valeur de λ choisie sera celle qui maximise le rapport entre la somme des carrés des écarts résiduels et la somme des carrés des écarts totale (SSE/SST)[147].

Les résultats obtenus sur base des différentes définitions possibles de λ ont été comparés par WU aux résultats obtenus par la méthode SAM. L'évaluation du taux de faux positifs (FDR) montre que l'utilisation de la version pénalisée du *t* de STUDENT produit moins de faux positifs que la méthode SAM, pour un même taux de vrais positifs (VP) détectés [147].

II.A.2.h. Autres corrections de la méthode SAM

BROBERG, en 2003, utilise la statistique d au coeur de SAM mais développe une méthode qui contrôle à la fois le nombre de gènes erronément déclarés faux positifs (FP) et le

nombre de gènes erronément déclarés faux négatifs (FN), en trouvant la valeur s_0 qui minimise la statistique *C*, définie par l'équation II.A.34 [27].

$$C = \sqrt{FP^2 + FN^2} \quad \text{(Equ. II.A.34)}$$

VAN DE WIEL (VAN DE WIEL, 2005) quant à lui remplace la statistique *t* modifiée proposée par TUSHER *ET AL.* par le test de la somme des rangs de WILCOXON et MANN & WHITNEY [137].

XIAO *ET AL.* (2002), appliquent une correction de WELCH à l'approche abordée dans SAM, conduisant à une relation plus instable entre les taux de faux positifs (FDR) et la proportion de gènes biologiquement impliqués détectés (sensibilité)[148]. En outre le choix de s_0 se base sur la minimisation du FDR plutôt que sur la minimisation du coefficient de variation de la statistique d (SAM) ou de la maximisation de la séparation des distributions (EBAM) [148].

En 2007, ZHANG rapporte également une étude comparative des implémentations de la méthode entre la version originale décrite par TUSHER, et la procédure implémentée dans la version 2.20 de SAM. La procédure de calcul du FDR dans la version 2.20 a été corrigée pour fournir des résultats plus corrects. La version originale surestime le FDR, car la distribution nulle des gènes différentiellement exprimés est plus dispersée que celle des gènes non impliqués. ZHANG soulève également un problème lié à la méthode de sélection, basée sur la distance entre le score observé et le score attendu, et une attribution du FDR qui repose sur une autre approche [152].

Dans certaines situations, l'utilisation de seuils asymétriques dans la version la plus récente produit des erreurs et des résultats contradictoires. ZHANG montre que les performances de SAM 2.20 (seuil symétrique), dépendent de la proportion de gènes induits et réprimés dans le jeu de données. La méthode du seuil de sélection symétrique fourni de meilleurs résultats lorsque les gènes induits et réprimés sont présents en même nombre. Cette avantage du seuil symétrique diminue graduellement, et SAM 2.20 peut éventuellement fournir de meilleurs résultats lorsque tous les gènes impliqués sont soit induits, soit réprimés [152].

II.A.2.i. La statistique B

LÖNNSTEDT & SPEED, en 2002, présentent une approche bayesienne différente, conduisant à la définition de la statistique B. Celle est définie par le rapport des probabilité d'expression différentielle et d'expression identiques (Equations II.A.35 et II.A.36) [100].

$$M_{ij}|\mu_i,\sigma_i \sim N(\mu_i,\sigma_i) \quad \text{(Equ. II.A.35)}$$

$$B_g = \log\frac{P(I_g=1|(M_{ij}))}{P(I_g=0|(M_{ij}))} \quad \text{(Equ. II.A.36)}$$

avec I_g la fonction qui indique si le gène est différentiellement exprimé ($I_g=1$) ou non ($I_g=0$), et M_{ij}, la moyenne des valeurs d'expression, distribuées suivant une normale.

L'application du théorème de BAYES conduit à l'expression de la statistique B sur base des valeurs d'expression observées, et la comparaison des probabilités obtenues si le gène est différentiellement exprimé ou s'il ne l'est pas. La participation de ces deux probabilités est modulée par la proportion relative des deux populations de gènes concernées (Equation II.A.37) [100].

$$B_g = \log\frac{p}{1-p}\frac{P((M_{ij})|I_g=1)}{P((M_{ij})|I_g=0)} = \log\frac{p}{1-p}\frac{P(M_g|I_g=1)}{P(M_g|I_g=0)} \quad \text{(Equ. II.A.37)}$$

avec M_g, le vecteur de valeurs d'expression individuelle.

Pour estimer ces probabilités *a priori*, l'ensemble des données est utilisé pour caractériser la moyenne et la variance sur base d'une distribution gamma pour $1/\sigma_i^2$ et d'une distribution normale pour μ_i avec une variance de valeur σ_i^2. Ces deux paramètres sont donc distribués conjointement (Equations II.A.38 et II.A.39) [100].

$$\tau_i = \frac{na}{2\sigma_i^2} \sim \Gamma(\nu,1) \quad \text{(Equ. II.A.38)}$$

$$\begin{aligned} \mu_i|\tau_i &= 0 && \text{if } I_i=0 \\ \mu_i|\tau_i &= N\left(0,\frac{cna}{2\tau_i}\right) && \text{if } I_i=1 \end{aligned} \quad \text{(Equ. II.A.39)}$$

avec $a>0$, le paramètre d'échelle associé à la variance, et $c>0$, le paramètre d'échelle qui exprime la dépendance entre μ_i et τ_i.

Le calcul des densités, et leur intégration, conduisent à deux statistiques *t* modifiées, et la statistique B_g est évaluée sur base de l'équation II.A.40.

$$B_g = \log\left(\frac{p}{1-p}\frac{1}{\sqrt{1+nc}}\left[\frac{a+s_g^2+M_g^2}{a+s_g^2+\frac{M_g^2}{1+nc}}\right]^{v+\frac{n}{2}}\right) \quad \text{(Equ. II.A.40)}$$

L'équation II.A.40 montre que le paramètre a joue le même rôle que le *fudge factor* de SAM, et assure que le rapport ne soit pas trop élevé lorsque la variance est petite et que le niveau d'expression est faible (M_g) [100, 136].

L'estimation des hyperparamètres p, v, a, c est réalisée sur base des données disponibles. Le paramètre p est fixé arbitrairement et sa valeur n'a pas beaucoup d'impact sur les résultats finaux. Les paramètres v et a sont extraits des variances observées par la méthode des moments. L'estimation de c est plus délicate car elle repose sur la connaissance de l'appartenance des gènes aux deux populations. Empiriquement, l'estimation de c est réalisée en prenant en compte les gènes les plus impliqués (proportion p) et ceux qui ne le sont pas (proportion $1-p$) [100].

II.A.2.j. Limma et le « moderated t »

Sur base d'un modèle bayesien hiérarchique similaire, SMYTH décrit le *moderated t*, implémenté dans le logiciel *Limma* [124].

Les mesures d'intensité fournies par la technologie des *microarrays* fournissent un vecteur de données d'expression $y_g^T = \left(y_{g1}, y_{g2}, \ldots y_{gn}\right)$, où chaque gène est indicé par la lettre g, pour chacune des n valeurs disponibles (n *microarrays*). En désignant par X la matrice de description du design de l'expérience, la valeur attendue de y_g peut être formulée par l'équation II.A.41, et sa variance est formulée par l'équation II.A.42 [124].

$$E(y_g) = X\alpha_g \quad \text{(Equ. II.A.41)}$$

$$var(Y_g)=W_g\sigma_g^2 \quad \text{(Equ. II.A.42)}$$

où α_g est un vecteur de coefficients associés à chaque gène, et W_g est une matrice de poids non négative. Les contrastes associés à certains coefficients individuels présentent un intérêt biologique. En désignant ceux-ci par β_g, l'hypothèse nulle du test peut s'écrire $\beta_{gj}=0$. L'utilisation d'un modèle linéaire pour modéliser les données d'expression peut être résumée par les équations II.A.43 à II.A.45 [124].

$$\beta_g=C^T\alpha_g \quad \text{(Equ. II.A.43)}$$

$$var(\alpha_g)=V_g s_g^2 \quad \text{(Equ. II.A.44)}$$

$$var(\hat{\beta}_g)=C^T V_g C s_g^2 \quad \text{(Equ. II.A.45)}$$

où s_g^2 est l'estimateur de σ_g^2, V_g est une matrice positive qui ne dépend pas de s_g^2. Les estimateurs de contrastes $\hat{\beta}_g$ suivent une distribution normale de moyenne β_g avec une matrice de covariance $C^T V_g C\sigma_g^2$ et s_g^2 suit une distribution χ^2 réduite, comme le montre les équations II.A.46 et II.A.47 [124].

$$\hat{\beta_{gj}}|\beta_{gj},\sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2) \quad \text{(Equ. II.A.46)}$$

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g} \quad \text{(Equ. II.A.47)}$$

où v_{gj} est l'élément à la position j de la diagonale de la matrice $C^T V_g C$, et d_g sont les degrés de liberté associés au gène g pour le modèle linéaire. La définition de la statistique *t* de STUDENT correspondante est formulée par l'équation II.A.48, et suit approximativement une distribution *t* avec d_g degrés de liberté [124].

$$t_{gj}=\frac{\hat{\beta_{gj}}}{s_g\sqrt{v_{gj}}} \quad \text{(Equ. II.A.48)}$$

Pour décrire la distribution de β_{gj} et de σ_g^2 pour tous les gènes, SMYTH décrit un modèle hiérarchique, tirant profit de la structure parallèle du jeu de données pour tous les gènes, et utilise le même modèle linéaire pour tous les gènes. L'estimation de ces paramètres repose

sur des *a priori* distributionnels. L'auteur considère pour tous les gènes que $\beta_{gj} \neq 0$ avec une probabilité égale à la proportion de gènes différentiellement p_j. La distribution de σ_g^2 repose sur la distribution d'un estimateur *a priori* s_0^2 qui suit une distribution χ^2 avec d_o degrés de liberté (Equ. II.A.49). Les informations *a priori* sur les coefficients est supposée suivre une distribution normale décrite par l'équation II.A.50 [124].

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0} \quad \text{(Equ. II.A.49)}$$

$$\beta_{gj} | \sigma_g^2, \beta \neq 0 \sim N(0, v_{0j} \sigma_g^2) \quad \text{(Equ. II.A.50)}$$

Ce modèle est similaire au modèle de LÖNNSTEDT & SPEED, avec $d_g = p$, $v_g = 1/n$, $d_0 = 2v$, $s_0^2 = a/(d_0 v_g)$ et $v_0 = c$ [100, 124]. Sur base du modèle hiérarchique, la moyenne *a posteriori* de σ_g^2, connaissant s_g^2, se calcule par l'équation II.A.51 et la statistique du *moderated t* est définie par l'équation II.A.52, et suit une distribution *t* avec $d_0 + d_g$ degrés de liberté [124].

$$\tilde{s}_g^2 = E(\sigma_g^2 | s_g^2) = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \quad \text{(Equ. II.A.51)}$$

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}} = \sqrt{\frac{d_0 + d_g}{d_g}} \frac{\hat{\beta}_{gj}}{\sqrt{s_{*g}^2 v_{gj}}} \quad \text{(Equ. II.A.52)}$$

$$s_{*g}^2 = s_g^2 + \frac{d_0}{d_g} s_0^2$$

L'intégration des fonctions de densité conduisent finalement les auteurs aux équations II.A.53 et II.A.54, qui définissent la distribution *t* suivie par la statistique du *moderated t* [124].

$$s^2 = s_0^2 F_{d, d_0} \quad \text{(Equ. II.A.53)}$$

$$\tilde{t} | \beta = 0 \sim t_{d_0 + d_g} \text{ et } \tilde{t} | \beta \neq 0 \sim \sqrt{1 + \frac{v_0}{v}} t_{d_0 + d_g} \quad \text{(Equ. II.A.54)}$$

Les hyperparamètres d_0 et s_0 qui interviennent dans le modèle sont estimés sur base des équations II.A.55 et II.A.56. Sur base d'une fonction cumulative qui est la somme des fonctions cumulatives des distribution t associées aux gènes identiquement et différentiellement exprimés, avec une proportion $1-p$ et p respectivement, le paramètre v_0 est estimé par l'équation II.A.57. Le paramètre p_j est fixé arbitrairement à 0.01, mais n'a pas une grande importance. Les fonctions ψ et ψ' sont les fonctions digamma et trigamma, respectivement [124].

$$\psi'\left(\frac{d_0}{2}\right)=mean\left\{(e_g-\bar{e})^2\frac{n}{n-1}-\psi'\left(\frac{d_g}{2}\right)\right\} \quad \text{(Equ. II.A.55)}$$

$$s_0^2=\exp\left\{\bar{e}+\psi\left(\frac{d_0}{2}\right)-\log\left(\frac{d_0}{2}\right)\right\} \quad \text{(Equ. II.A.56)}$$

$$v_0=v_g\left(\frac{\tilde{t}_g^2}{q_{target}^2}-1\right)$$

$$q_{target}=F^{-1}(p_{target};d_0+d_g) \qquad \text{(Equ. II.A.57)}$$

$$p_{target}=\frac{1}{p}\left\{\frac{r-0.5}{2\mathrm{G}}-(1-p)F(-|\tilde{t}_g|;d_0+d_g)\right\}$$

où r est le rang du gène, et G est le nombre total de gènes. En pratique, v_0 est calculé pour les valeurs de r allant de 1 à $Gp/2$, et la valeur finale de v_0 utilisée est la moyenne de ces valeurs.

Afin de montrer la corrélation du *moderated t* avec la statistique B présentée au paragraphe précédent, nous pouvons reformuler la définition des probabilités étudiées similairement avec la statistique B (Equation II.A.58) [124].

$$B_{gj}=\log O_{gj}$$

$$O_{gj}=\frac{p_j}{1-p_j}\frac{p(\tilde{t}_{gj}|\beta_{gj}\neq 0)}{p(\tilde{t}_{gj}|\beta_{gj}=0)}=\frac{p_j}{1-p_j}\sqrt{\frac{v_{gj}}{v_{gj}-v_{0\mathrm{j}}}}\left(\frac{\tilde{t}_{gj}^2+d_0+d_g}{\tilde{t}_{gj}^2\frac{v_{gj}}{v_{gj}+v_{0\mathrm{j}}}+d_0+d_g}\right)^{\frac{1+d_0+d_g}{2}} \quad \text{(Equ. II.A.58)}$$

II.A.2.k. Le shrinkage-t : utilisation d'un estimateur de type Stein

OPGEN-RHEIN & STRIMMER, en 2007, proposent, à l'instar de CUI *ET AL.* en 2005, d'utiliser l'estimation décrite par JAMES-STEIN [38, 109]. La méthode ne repose donc pas sur des postulats distributionnels, fournissant une statistique individuelle utilisable telle quelle pour caractériser les gènes [109].

En désignant par $\hat{V}$ le vecteur des variances individuelles observées, la règle de JAMES-STEIN est formulée par l'équation II.A.59.

$$\delta^{\lambda}=\delta^{0}-\lambda\Delta=\hat{V}-\lambda(\hat{V}-\hat{V}^{Target})=\lambda\hat{V}^{Target}+(1-\lambda)\hat{V} \quad \text{(Equ. II.A.59)}$$

Le paramètre λ détermine le poids de chacun des deux estimateurs combinés linéairement pour calculer l'estimateur *shrinkage*, δ^{λ}. La méthodologie repose sur une estimation optimale de λ. La règle suivie par les auteurs vise à minimiser l'erreur commise sur l'estimation de δ^{λ}, en comparant le carré de l'erreur observée avec le carré de l'erreur attendue (le carré moyen calculé sur base des données). La formulation mathématique de cette règle est décrite par l'équation II.A.60, qui montre que λ peut-être déterminé sans référence à la valeur réelle de V. La définition mathématique de cette règle fournit une parabole définie par les paramètres a, b et c, déterminés sur base des deux premiers moments distributionnels de $\hat{V}$ et de $\hat{V}_{target}$ [109].

$$MSE(\delta^{\lambda})=c+\lambda^{2}b-2\lambda a$$

$$c=MSE(\hat{V})$$

$$b=\sum_{k=1}^{p}\left\{E\left((\hat{V}_{k}-\hat{V}_{k}^{Target})^{2}\right)\right\} \quad \text{(Equ. II.A.60)}$$

$$a=\sum_{k=1}^{p}\left\{Var(\hat{V}_{k})-Cov(\hat{V}_{k},\hat{V}_{k}^{Target})+Bias(\hat{V}_{k})\,E(\hat{V}_{k}-\hat{V}_{k}^{Target})\right\}$$

Lorsque $\lambda=0, MSE\left(\delta^{\lambda}=\delta^{0}\right)=MSE(\hat{V})$, et lorsque l'estimateur *shrinkage* est utilisé, la correction apportée repose uniquement sur les valeurs de a et de b. Toute valeur de λ comprise entre 0 et $2a/b$ implique une diminution de la valeur de $MSE\left(\delta^{\lambda}\right)$. La valeur

optimale de λ est le minimum de la parabole, dont les coordonnées sont $\lambda^* = a/b$ et $MSE(\hat{V}) - MSE\left(\delta^{\lambda^*}\right) = a2/b$. Les deux valeurs de a et b, ou alternativement le rapport a/b, peuvent être estimées empiriquement [109].

Sur base du modèle de LINDLEY & SMITH (1972, [97]), en n'utilisant que la partie positive de l'estimateur $min(1,\lambda)$, et en évitant un correction trop importante, les auteurs reformulent le test suivant l'équation II.A.61. La valeur M est un seuil défini par l'utilisateur, qui correspond à la différence maximale autorisée entre l'estimateur *shrinkage* et l'estimateur classique. Le choix de la valeur minimum entre 1 et $M/|\Delta_k|$ définit l'importance de la correction apportée [109].

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^0 - min(1,\hat{\lambda})\, min\left(1,\frac{M}{|\Delta_k|}\right)\Delta_k \quad \text{(Equ. II.A.61).}$$

En accord avec les autres méthodes présentées, les auteurs utilisent la statistique *t* classique sur base d'un estimateur corrigé de la variance, suivant le modèle de type JAMES-STEIN (Equations II.A.59 et II.A.61). La variance individuelle est désignée par v_k. Les auteurs définissent la valeur de v^{Target} sur base de la valeur médiane de v_k, et la covariance entre v_k et v_{median} est considérée comme nulle. Le numérateur assure que la correction est apportée lorsque la dispersion de la variance est importante, et le dénominateur évite une correction trop importante lorsque l'estimateur v_{median} est un mauvais choix (Equation II.A.62) [109].

$$v_k{}^* = \hat{\lambda}^*\, v_{median} + (1-\hat{\lambda}^*)\, v_k$$

$$\hat{\lambda}^* = min\left(1, \frac{\sum_{k=1}^{p} \widehat{Var}(v_k)}{\sum_{k=1}^{p} (v_k - v_{median})^2}\right) \quad \text{(Equ. II.A.62)}$$

La statistique *t* obtenue en utilisant l'estimateur *shrinkage* est utilisée telle quelle pour sélectionner les gènes différentiellement exprimés, sans lui assigner de significativité en raison de l'absence de conditions distributionnelles [109].

Bien que le modèle utilisé soit plus simple que le modèle bayesien complet décrit par

Smyth, les performances de la statistique *moderated t* et de la statistique *shrinkage t* sont similaires et fournissent les meilleurs résultats, sur base des simulations réalisées par Opgen-Rhein & Strimmer [18, 109].

II.B.
Analyse de groupes

Résumé

Ce chapitre présente un échantillon non exhaustif des méthodes actuelles d'analyse de l'expression de groupe de gènes. Les méthodes sélectionnées fournissent néanmoins un aperçu représentatif des différentes stratégies publiées. Par « analyse de groupes », nous désignons les études qui s'intéressent à la manière dont un groupe fonctionnel connu évolue avec les conditions de l'expérience. Ceci les distingue des études de coexpression qui recherchent des corrélations dans l'expression de gènes, et des méthodes dites de *clustering*, qui visent à découvrir de nouveau groupes (*group discovery*) sur base des conditions de l'expérience.

L'examen des différentes procédures disponibles actuellement révèle trois catégories de méthodes d'analyse de groupes :

- ☞ Les méthodes de sur-représentation, qui étudient les groupes sur base du nombre de gènes dont la modification d'expression est la plus significative, et qui ne considèrent que les gènes les plus significatifs ;
- ☞ Les méthodes post-hoc, qui utilisent la totalité des résultats de l'analyse individuelle pour en dériver une statistique secondaire associée au groupe, soit en utilisant un score d'enrichissement évalué lors du parcours de la liste des gènes, soit en calculant une statistique qui résume la réponse des membres du groupe ;
- ☞ Les méthodes globales, apparues récemment, suggèrent d'utiliser un modèle multivarié pour décrire les données d'expression observées et en déduire la significativité.

Ainsi, chronologiquement, les sujets liés à l'analyse de groupe ont évolué, depuis les questions du types « quel groupe est plus représenté que les autres au sein de la liste des gènes les plus significatifs ? » vers la question « quels sont les groupes au sein desquels l'expression est globalement différente entre les conditions ? » Nous montrerons, dans la troisième partie du chapitre Résultats, comment répondre avec *FAERI* à la question « quels sont les groupes dont les membres, ou une partie d'entre eux, se comportent différemment entre les conditions, peu importe la direction ? »

Nous portons également une attention particulière sur la classification des méthodes de groupes, selon la cohérence entre la définition de l'hypothèse et la procédure utilisée pour évaluer la significativité. Nous pouvons remarquer trois types de *scenarii*, parmi les méthodes post-hoc et globales : une hypothèse nulle compétitive évaluée avec une procédure correcte, avec une procédure auto-suffisante, ou encore une hypothèse nulle auto-suffisante évaluée correctement.

II.B.1. Introduction

Dans le contexte de la biologie moléculaire, l'analyse individuelle de l'expression différentielle des gènes permet de mettre en évidence l'implication de différents gènes participant de manière significative au sujet d'étude envisagé. Toutefois, une telle analyse se doit d'être complétée par une caractérisation systémique des mécanismes moléculaires impliqués. Sur base des résultats obtenus grâce aux méthodes d'analyse individuelles, il est possible de mettre en évidence certains mécanismes régulateurs ou cibles d'une pathologie ou d'un médicament étudié. Ainsi, les recherches situées en aval de l'analyse de l'expression différentielle des gènes permettent, sur base de voies métaboliques connues, ou d'autres critères biologiques (localisation chromosomique, facteurs de transcription impliqués...) de « grouper » les gènes de façon pertinente. Lorsque plusieurs gènes, partageant une caractéristique commune, se situent dans la liste des gènes les plus significativement impliqués, nous pouvons supposer que ce critère est impliqué dans la modification du profil d'expression, et donc dans la pathologie étudiée.

Cependant, selon la méthode utilisée et le seuil de sélection choisi par le chercheur, ce relevé systématique peut s'avérer être un vrai travail de fourmis. D'une part, lorsque plusieurs centaines de gènes sont considérés comme significatifs, leur regroupement sur base de différents critères est un travail de longue haleine. Cette démarche vise à « donner un sens biologique » à une liste de gènes. Selon le critère de regroupement utilisé, l'interprétation de la liste de gènes obtenue peut être différente. Plusieurs méthodes d'analyses, situées en aval de la sélection des gènes impliqués, permettent de quantifier et d'étudier la significativité des groupes de gènes envisagés.

Cette manière de procéder s'avère incomplète, pour deux raisons majeures. D'une part, elle repose sur une sélection initiale, dont le seuil de détection est arbitrairement choisi par l'analyste. Selon le seuil utilisé, le nombre de gènes considérés sera différent. Par conséquent, la représentation des différentes voies métaboliques (ou d'un autre critère) au sein de la *top-list* en sera affectée, et les conclusions de l'expérience seront différentes, incluant un nombre variable de groupes de gènes et induisant une vision systémique différente de la même problématique. L'origine de cette erreur tient essentiellement de la transformation d'une liste de *p-values* individuelles (distribution continue) en une classification binaire séparant les gènes « significatifs » des gènes « non significatifs » (valeurs binaires 0/1 ou Vrai/Faux).

D'autre part, d'un point de vue biologique, un groupe de gènes pour lesquels une faible variation est enregistrée sur tous les gènes peut avoir un impact régulateur bien plus important qu'un groupe de gènes caractérisé par une variation importante du niveau d'expression de quelques-uns de ses membres. Ainsi, sur base de cet exemple, le premier cas de figure sera complètement ignoré par une méthode d'analyse placée en aval de la détection des gènes, malgré son importance phénotypique, simplement parce qu'aucun des membres du groupe n'apparaît significatif. Ce cas de figure, envisagé par un second groupe de méthodologistes, a motivé la création de nouvelles méthodes d'analyses, où les résultats de l'analyse individuelle sont utilisés dans leur intégralité pour quantifier la réponse de tous les groupes de gènes représentés par le jeu de données, quelle que soit la magnitude de la réponse individuelle. Ces méthodes se situent donc en aval de l'étude individuelle de l'expression différentielle, mais n'implique aucune sélection préalable des gènes les plus impliqués. La significativité des groupes y est calculée sur base de la liste des *p-values* individuelles, du rang des gènes dans cette listes, ou de la statistique utilisée pour dresser cette liste (valeur *t*, *fold change*...).

Cette première distinction entre les deux approches analytiques envisagées ouvre la porte à une troisième catégorie de méthodes, où l'implication des groupes de gènes est étudiée au départ des données brutes, indépendamment des résultats obtenus avec une méthode d'analyse individuelle.

Au fil des années, les deux dernières approches tendent à se généraliser, la stratégie d'analyse repose de plus en plus sur une analyse des groupes de gènes sur base de l'ensemble des données, affinée par les méthodes d'analyse individuelles qui visent à clarifier, au sein des différents groupes de gènes, la participation individuelle de chaque membre.

Les différentes méthodes d'analyse de groupes de gènes peuvent également être classées sur base d'autres critères. En effet, d'un point de vue méthodologique, chaque méthode repose sur une série d'hypothèses, sur le choix d'une statistique représentative et sur la procédure utilisée pour estimer la significativité des groupes de gènes. Dans les paragraphes suivants, chacune des méthodes existante est détaillée, en veillant à accorder une attention particulière à ces choix et à leurs implications, telles que suggérées par Goeman & Bühlmann en 2007. Suivant leurs arguments, les méthodes seront dites « compétitives » lorsque la significativité d'un groupe de gènes est estimée par rapport aux résultats des autres groupes, ou « auto-suffisantes » (*self-contained*), lorsque seules les données du groupe étudié sont utilisées pour attribuer la significativité. D'autre part, la

significativité étant fréquemment calculée sur base d'une distribution nulle obtenue par permutations, une attention particulière sera portée sur l'interprétation des résultats. En effet, la genèse des permutations peut être réalisée sur base d'un ré-échantillonnage aléatoire des gènes dans un groupe (*gene-sampling*) ou d'un échantillonnage de labels (phénotype) des valeurs individuelles (démarche statistique traditionnelle). En particulier, cette classification permet de mettre en évidence l'existence de méthodes « hybrides » (GSEA et ses dérivés) où l'hypothèse nulle repose sur une échantillonnage des gènes mais où la significativité est attribuée sur base d'un échantillonnage de labels, générant des résultats qu'il faut interpréter avec précaution.

Enfin, en plus de la classification proposée, nous porterons particulièrement l'attention du lecteur sur d'autres critères biologiques, tel que la capacité des méthodes à détecter des groupes de gènes dont les membres fournissent une réponse différente. Un groupe de gène pour lequel 50 % des membres sont sur-exprimés, et 50% sont sous-exprimé peut, selon les cas, fournir un résultat « moyen » ou « additif », conduisant à le considérer comme non significatif, ou significatif, respectivement.

II.B.2. Les méthodes de sur-représentation

Les premières études portant leur intérêt sur l'expression différentielle observée au sein de groupes de gènes ont été rapportées au début de cette décennie [32, 45, 79, 82-85]. Ces études pionnières se basent sur la comparaison croisée de la liste des gènes détectés comme différentiellement exprimés, avec la liste des gènes appartenant à un même groupe. L'approche utilisée dans ces études présente plusieurs points communs. Dans un premier temps, les résultats de l'analyse individuelle des gènes sont utilisés, en conjonction avec un seuil de sélection arbitrairement choisi (*p-value*) pour définir une liste de gènes considérés comme différentiellement exprimés. D'un point de vue informatique, cette étape consiste à créer une liste binaire (valeurs 0/1 ou Non Détecté / Détecté) au départ de la distribution observée de la statistique étudiée (*fold change*, valeur *t*, *p-value*, ...). La seconde étape repose sur la comparaison de cette liste binaire avec une seconde liste binaire, caractérisant l'appartenance du gène au sein du groupe de gènes étudié (voie métabolique, groupe fonctionnel, ...).

Cette comparaison est illustrée classiquement à l'aide de tables 2*2, dont la formulation générale est représentée dans le tableau II.B.1. Chaque catégorie de la table 2*2 correspond au nombre de gènes répondant à l'une des définitions suivantes : (i) le gène est membre du groupe et est détecté, (ii) il est membre du groupe mais n'est pas détecté, (iii) il est détecté mais n'appartient pas groupe, ou (iv) il n'est ni détecté, ni membre du groupe de gènes [32, 45, 79, 82-85].

	Sélectionné	**Non-Sélectionné**	**Total**
Membre du groupe	n_{SG}	n_{oG}	n_{*G}
Non Membre	n_{So}	n_{oo}	n_{*o}
Total	n_{S*}	n_{o*}	n_{**}

Table II.B.1 : Représentation générale d'une table 2*2 utilisée pour caractériser un groupe de gènes sur base des résultats de l'analyse individuelle.

La seconde étape inhérente aux tests 2*2 repose sur l'utilisation de cette table pour calculer une statistique caractérisant le groupe de gènes étudié. Les modèles statistiques employés reposent sur les modèles hypergéométrique, binomial, khi-carré (χ^2) ou le test exact de FISHER [82-85]. La significativité du groupe est ensuite évaluée sur base de différentes distributions nulles, qui diffèrent selon les auteurs. Elle peut être évaluée sur base d'un

modèle théorique, ou sur base de permutations où la liste d'appartenance au groupe est générée de manière aléatoire, avec ou sans remplacement (permutations relatives aux gènes). Les divergences entre les études reposent sur la statistique employée et sur la procédure de détermination de la distribution nulle, en conformité avec le modèle choisi.

Selon le modèle hypergéométrique, la probabilité qu'un groupe de gènes soit représenté par n_{SG} membres au sein de la liste des gènes sélectionnés (n_{S*}) peut se calculer en générant aléatoirement des groupes de gènes de taille n_{*G}, sur base de l'équation II.B.1.

$$P(n_{SG}|n_{**}, n_{*G}, n_{S*}) = \frac{\binom{n_{*G}}{n_{SG}}\binom{n_{**}-n_{*G}}{n_{S*}-n_{SG}}}{\binom{n_{**}}{n_{S*}}} = \frac{\binom{n_{*G}}{n_{SG}}\binom{n_{*0}}{n_{S0}}}{\binom{n_{**}}{n_{S*}}} \quad \text{(Equ. II.B.1)}$$

La probabilité d'avoir n_{SG} gènes ou moins dans le groupe se calcule alors par la somme des probabilités d'obtenir 1, 2 ... n_{SG} membres du groupe dans une liste aléatoire de n_{S*} gènes (Equation II.B.2).

$$p = \sum_{i=0}^{n_{SG}} \frac{\binom{n_{*G}}{i}\binom{n_{**}-n_{*G}}{n_{S*}-i}}{\binom{n_{**}}{n_{S*}}} \quad \text{(Equ. II.B.2)}$$

Et la *p-value* pour les catégories sur-représentées se calcule par l'équation II.B.3.

$$p = 1 - \sum_{i=0}^{n_{SG}} \frac{\binom{n_{*G}}{i}\binom{n_{**}-n_{*G}}{n_{S*}-i}}{\binom{n_{**}}{n_{S*}}} \quad \text{(Equ. II.B.3)}$$

Lorsque n_{**} est élevé, la distribution hypergéométrique tend vers une distribution binomiale, et les équations II.B.1 et II.B.3 peuvent être remplacées par les équations II.B.4 et II.B.5.

$$P\left(n_{SG}|n_{S*},\frac{n_{*G}}{n_{**}}\right)=\binom{n_{S*}}{x}\left(\frac{n_{*G}}{n_{**}}\right)^{x}\left(1-\frac{n_{*G}}{n_{**}}\right)^{n_{S*}-x} \quad \text{(Equ. II.B.4)}$$

$$p=1-\sum_{i=0}^{n_{SG}-1}\binom{n_{S*}}{i}\left(\frac{n_{*G}}{n_{**}}\right)^{i}\left(1-\frac{n_{*G}}{n_{**}}\right)^{n_{S*}-i} \quad \text{(Equ. II.B.5)}$$

Le test pour l'égalité de deux proportions, ou test du χ^2 (Khi-carré), implique le calcul de la valeur observée du χ^2 (Equation II.B.6), et sa comparaison avec la distribution théorique du χ^2 avec 1 degré de liberté ($df=(2-1)*(2-1)=1$) [45, 57, 101].

$$\chi^2=\frac{n_{**}\left(\left|n_{0G}n_{S0}-n_{SG}n_{00}\right|-\frac{n_{**}}{2}\right)^2}{n_{*G}n_{*0}n_{0*}n_{S*}} \quad \text{(Equ. II.B.6)}$$

Lorsque le nombre d'échantillons est élevé, le terme de correction de continuité du numérateur ($n_{**}/2$) peut être omis.

La relation présentée dans l'équation IV.B.6 est équivalente à la statistique Z utilisée par KAL *ET AL.*, car il a été démontré que $\chi^2=Z^2$ (Equation II.B.7) [57, 79].

$$Z=\frac{p_{0G}-p_{SG}}{\sqrt{p(1-p)\left(\frac{1}{n_{0*}}+\frac{1}{n_{S*}}\right)}} \text{ avec } p_{0G}=\frac{n_{0G}}{n_{0*}}, p_{SG}=\frac{n_{SG}}{n_{S*}}, p=\frac{p_{0G}+p_{SG}}{2} \quad \text{(Equ. II.B.7)}$$

Si le nombre de valeurs dans une catégorie de la table 2*2 est insuffisant (moins de 5 valeurs), le test du χ^2 devient inapproprié. Dans ce cas, le test hypergéométrique, ou test exact de FISHER, peut être utilisé pour calculer la probabilité d'obtenir chaque configuration possible de la table 2*2, sur base des valeurs totales de chaque catégorie (Equation II.B.8). La *p-value* est ensuite calculée en listant l'ensemble des tables de 2*2 qui fournissent au minimum la probabilité obtenue [45, 57, 101, 127].

$$P=\frac{n_{S*}!n_{0*}!n_{*G}!n_{*0}!}{n_{**}!n_{SG}!n_{S0}!n_{0G}!n_{00}!} \quad \text{(Equ. II.B.8)}$$

Enfin, AUDIC & CLAVERIE ont publié une étude rapportant l'utilisation d'une distribution de

POISSON, sur base d'une approche bayesienne. Ils étudient la probabilité d'observer le nombre de gènes sélectionnés et membres du groupe de gènes (n_{SG}) sur base du nombre de gènes non sélectionnés et membres du groupe de gènes (n_{0G}) (Equations II.B.9 et II.B.10) [10].

$$P(n_{SG}|n_{0G}) = \left(\frac{n_{S*}}{n_{0*}}\right)^{n_{SG}} \frac{(n_{0G}+n_{SG})!}{n_{0G}!n_{SG}!\left(1+\frac{n_{S*}}{n_{0*}}\right)^{(n_{0G}+n_{SG}+1)}} \quad \text{(Equ. II.B.9)}$$

$$p = min\left\{\sum_{k=0}^{k\leq n_{SG}} P(k|n_{0G}), \sum_{k=n_{SG}}^{\infty} P(k|n_{0G})\right\} \quad \text{(Equ. II.B.10)}$$

II.B.3. Les méthodes post-hoc de parcours de la liste des gènes

Plusieurs auteurs ont proposé des méthodes alternatives d'analyse de groupes de gènes. Motivés par leur désaccord avec l'utilisation d'un seuil de sélection strict sur les résultats de l'analyse individuelle des gènes, leurs méthodes utilisent soit la liste des statistiques individuelles, soit la liste des *p-values* associées aux tests individuels.

BREITLING *ET AL.*, en 2004, et AL-SHAHROUR *ET AL.*, en 2005, utilisent également les tables 2*2, mais leurs méthodes reposent sur une analyse simultanée des groupes de gènes pour plusieurs seuils de sélection sur les résultats de l'analyse individuelle [25, 26]. VIRTANEVA *ET AL.*, en 2001, utilisent quant à eux la statistique de la somme des rangs de WILCOXON comme statistique globale [140].

II.B.3.a. La procédure GSEA originale

L'une des méthodes les plus employée actuellement, dénommée GSEA (*Gene Set Enrichment Analysis*), a été décrite pour la première fois en 2003 par MOOTHA *ET AL.*, et repose sur un test d'enrichissement dérivé du test de KOLMOGOROV-SMIRNOV (*KS-test*). Dans sa forme originale, la procédure de GSEA évalue si les scores attribués aux gènes du groupe suivent une distribution uniforme. Le score utilisé pour chaque gène est représentatif de la différence observée entre les deux conditions comparées, par exemple le rapport entre le signal et le bruit ($SNR=(m_1-m_2)/(s_1+s_2)$). Dans un premier temps, les différents gènes sont ordonnés sur base de ce score, et un rang leur est attribué. Pour chaque groupe de gènes étudié, la procédure calcule un score d'enrichissement (ES = Statistique KS normalisée), définit par les équations II.B.11 et II.B.12 [106].

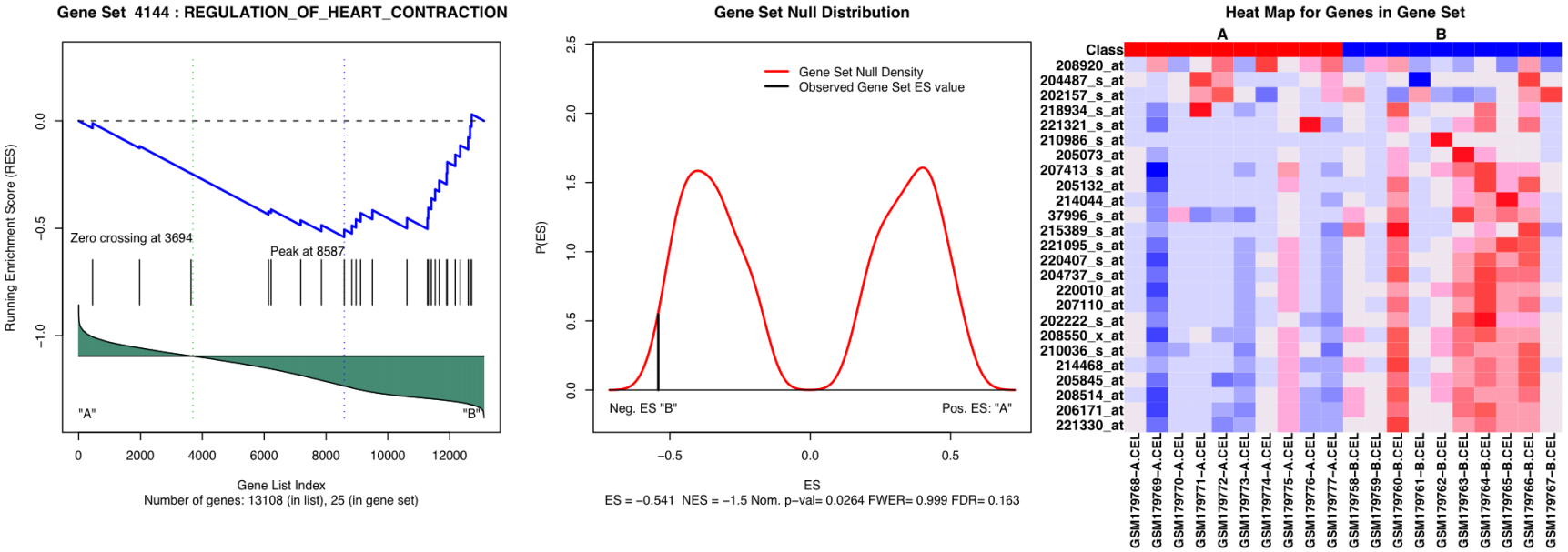


Figure II.B.1 : Exemple de graphique généré par la méthodologie GSEA. Le graphique de gauche illustre l'évolution du score ES lors du parcours de la liste des gènes. La valeur ES affectée au groupe de gènes étudiés correspond au minimum de la courbe observée. La graphique central illustre la distribution des scores obtenus pour des permutations de labels, et indique la valeur observée pour le groupe analysé. Le graphique de droite illustre les données d'expression de chaque *probeset* membre du groupe. Le jeu de données utilisé est E-GEOD-7429.

$$X_{i \notin G} = -\sqrt{\frac{n_{*G}}{n_{**} - n_{*G}}} \text{ et } X_{i \in G} = \sqrt{\frac{n_{**} - n_{*G}}{n_{*G}}} \quad \text{(Equ. II.B.11)}$$

$$MES = \max_{1 \leq j \leq n**} (ES) = \max_{1 \leq j \leq n**} \left(\sum_{i=1}^{j} X_i \right) \quad \text{(Equ. II.B.12)}$$

La procédure consiste, pour chaque groupe de gènes, à calculer la statistique ES en parcourant la liste de gènes, de la statistique la plus élevée à la statistique la plus faible (le SNR dans la version originale de GSEA). Un graphique illustrant ce parcours est représenté dans la figure II.B.1, illustrant l'évolution du score d'enrichissement au cours de la progression au sein de la liste de gènes. La valeur finale utilisée pour caractériser statistiquement le groupe de gène correspond à la valeur maximale atteinte par la valeur ES au cours de ce parcours [106].

En reformulant les équations II.B.11 et II.B.12, il apparaît que la fonction ES, qui calcule le score d'enrichissement, est le résultat de la différence de deux fonctions cumulatives distinctes ES_{*G} et E_{*0}, calculées sur base des équations II.B.11 et II.B.13 [106].

$$MES = \max_{1 \leq j \leq n**} \left(ES_{*G} - ES_{*0} \right) = \max_{1 \leq j \leq n**} \left(\sum_{i=0}^{j} X_{i \in G} - \sum_{i=0}^{j} X_{i \notin G} \right) \quad \text{(Equ. II.B.13)}$$

Cette procédure peut être qualifiée de « compétitive » car elle est basée sur la comparaison du groupe de gène avec le reste du jeu de données (son complément). L'hypothèse nulle qui est testée est « le groupe de gènes n'est pas associé au conditions comparées », et l'attribution de la *p-value* du score d'enrichissement repose sur des permutations d'échantillons, alors que les méthodes compétitives utilisent des permutations de gènes. Cette particularité fait de GSEA une méthode hybride dont les résultats sont particulièrement difficiles à interpréter.

II.B.3.b. La procédure définitive de GSEA

La procédure GSEA a ensuite été révisée et généralisée par SUBRAMANIAN en 2005. Dans cette nouvelle version de GSEA, la corrélation entre chaque gène et le phénotype (r_i) est utilisée pour pondérer l'incrément réalisé pendant le parcours de la liste. L' équation II.B.14 définit les statistiques utilisées, par comparaison avec l'équation II.B.13. Lorsque

$p=0$, le score d'enrichissement correspond à la statistique standard de KOLMOGOROV-SMIRNOV (KS), et lorsque $p=1$, la procédure utilise la corrélation individuelle pour pondérer le parcours [131].

$$\begin{aligned} MES &= \max_{1\le j\le n**}\left(ES_{*G}-ES_{*0}\right) \\ &= \max_{1\le j\le n**}\left(\sum_{i=1}^{j} X_{i\in G}-\sum_{i=1}^{j} X_{i\notin G}\right) \\ &= \max_{1\le j\le n**}\left(\sum_{i=1}^{j}\frac{\left|r_{i\in G}\right|^{p}}{\sum_{i\in G}\left|r_i\right|^{p}}-\sum_{i=1}^{j}\frac{1}{n_{*0}}\right) \end{aligned} \quad \text{(Equ. II.B.14)}$$

L'utilisation d'un parcours impliquant des poids individuels conduit à une asymétrie de la distribution des scores ES lorsque la majorité des gènes est corrélée avec l'un des deux phénotypes comparés. Par conséquent, l'estimation de la significativité doit être réalisée séparément pour les groupes de gènes présentant un score positif et un score négatif [131].

D'autre part, dans les améliorations décrites par SUBRAMANIAN, figure également une procédure de sélection basée sur le FDR. La *p-value* nominale y est toujours calculée sur base de permutations des échantillons. La correction pour tests multiples repose sur une normalisation de la valeur ES de chaque groupe de gènes, tenant compte de la taille du groupe (NES = Score d'enrichissement normalisé – Equation II.B.15). Le FDR est ensuite calculé en comparant la distribution observée et la distribution nulle de la statistique normalisée (NES)[131].

$$\begin{aligned} MES &= \max_{1\le j\le n**}\left(ES_{*G}-ES_{*0}\right) \\ &= \max_{1\le j\le n**}\left(\sum_{i=1}^{j} X_{i\in G}-\sum_{i=1}^{j} X_{i\notin G}\right) \\ &= \max_{1\le j\le n**}\left(\sum_{i=1}^{j}\frac{\left|r_{i\in G}\right|^{p}}{\sum_{i\in G}\left|r_i\right|^{p}}-\sum_{i=1}^{j}\frac{1}{n_{*0}}\right) \end{aligned} \quad \text{(Equ. II.B.15)}$$

Etant donné que la distribution nulle est utilisée pour assigner une *p-value* au score d'enrichissement, et que celle-ci est obtenue par permutations, plusieurs remarques doivent être formulées :

- ☞ la distribution des *p-values* obtenues suit une distribution discrète, présentant $(n_{perm}+1)$ valeurs à intervalles réguliers de taille $1/n_{perm}$;
- ☞ les permutations de label des échantillons sur un petit jeu de données ne permettent pas d'obtenir une très grande résolution dans la distribution finale des *p-values*. Plus le nombre de permutations possible est petit, plus grand sont les intervalles, et par conséquent plus il y a de groupes de gènes partageant le même rang dans la liste de *p-values* ;
- ☞ pour les grands jeux de données, un plus grand nombre de permutations est accessible, mais le temps de calcul de la distribution nulle limite la résolution accessible dans un délai raisonnable.

II.B.3.c. Adaptation mathématique de la procédure GSEA

Pour ces raisons, KELLER *ET AL.* ont décrit une procédure similaire à GSEA en adaptant leur algorithme de manière à pouvoir calculer une distribution nulle sur base mathématique [80]. L'évolution du score d'enrichissement lors du parcours de la liste de gènes se fait par paliers réguliers, sans pondération, de telle sorte que la valeur finale soit nulle (Equation II.B.16).

$$\begin{aligned} MES &= \max_{1\le j\le n_{**}} \left(ES_{*G} - ES_{*0} \right) \\ &= \max_{1\le j\le n_{**}} \left(\sum_{i=1}^{j} X_{i\in G} - \sum_{i=1}^{j} X_{i\notin G} \right) \\ &= \max_{1\le j\le n_{**}} \left(\sum_{i=1}^{j} \left(n_{*0} \right)_{i\in G} - \sum_{i=1}^{j} \left(n_{*G} \right)_{i\notin G} \right) \end{aligned} \quad \text{(Equ. II.B.16)}$$

Ainsi, la valeur théorique de la statistique MES atteint une valeur maximale égale à $\left(n_{*0} n_{*G} \right)$ et une valeur minimale égale à $-(n_{*0}\, n_{*G})$. De plus, la distribution des valeurs MES possibles peut être calculée sur base de la liste de l'ensemble des possibilités accessibles. L'équation II.B.17 permet de calculer le nombre de combinaisons possibles sur base de la distribution binomiale [80].

$$n_{ES} = \binom{n_{**}}{n_{*G}} = 2 n_{**} n_{*0} + 1 \quad \text{(Equ. II.B.17)}$$

La *p-value* est calculée par la probabilité qu'un score d'enrichissement aléatoire atteigne ou dépasse la valeur MES. Cette probabilité est calculée sur base de l'événement complémentaire (la probabilité que le score d'enrichissement aie une valeur inférieure à MES)[80].

La procédure de KELLER consiste ensuite à utiliser une fonction mathématique récursive pour remplir une matrice ou chaque ligne correspond à une valeur possible du score d'enrichissement et chaque colonne correspond à une étape du parcours dans la liste des gènes. Chaque élément (x,y) de la matrice se voit attribuer le nombre de possibilités d'obtenir le score d'enrichissement x à l'étape y, dont la valeur est comprise dans l'intervalle $]-|MES|;|MES|[$ [80].

Il est important de constater que pour remplir ce tableau, il n'est nécessaire de calculer que les valeurs d'enrichissement comprise dans l'intervalle défini. Plus grande est la valeur $|MES|$, plus le nombre de valeurs comprises dans l'intervalle est élevé, et plus la *p-value* sera petite. Le temps de calcul de la *p-value* est plus élevé pour les groupes de gènes présentant un score $|MES|$ plus grand [80].

II.B.4. Les méthodes post-hoc « auto-suffisantes »

Plusieurs méthodes d'analyse de groupes de gènes reposent sur le calcul d'une statistique unique au départ des résultats individuels. A la différence de GSEA et de ses dérivés, le calcul d'une statistique « auto-suffisante » ne repose pas sur l'évolution d'un score d'enrichissement lors du parcours de la liste des gènes. A l'inverse, ces méthodes ne tiennent compte que des statistiques individuelles associées aux membres du groupes, et en dérivent une statistique unique, attribuée au groupe de gènes.

II.B.4.a. Utilisation de la p-value individuelle

PAVLIDIS *ET AL.* (2004), utilisent une méthode simple pour assigner un score à chaque groupe de gènes. La première étape consiste à caractériser l'analyse individuelle à l'aide de coefficients de corrélation de PEARSON, fournissant une *p-value* pour chaque gène sur base d'une distribution normale. Ensuite, ils utilisent la moyenne arithmétique du logarithme des *p-values* individuelles (Equation II.B.18) pour définir le score du groupe. La significativité du groupe est ensuite calculée sur base du modèle compétitif, mettant en oeuvre des permutations de gènes [114].

$$Exp.\ Score = \frac{\sum_{i=0}^{n_{*G}} -\log(p_i)}{n_{*G}} \quad \text{(Equ. II.B.18)}$$

La distribution nulle étant différente pour tous les groupes ayant une taille différente, celle-ci doit être calculée pour chaque taille de groupe représentée dans la banque de donnée des groupes de gènes.

II.B.4.b. Le théorème central limite, le fold change, et la statistique Z

KIM & VOLSKY ont décrit en 2005 un méthode paramétrique d'analyse de groupes de gènes en s'appuyant sur le théorème central limite (PAGE, *Parametric Analysis of Gene Set Enrichment*) [87].

Selon ce théorème, la moyenne estimée au départ d'un échantillonnage aléatoire suit une distribution normale dont la moyenne est identique à la moyenne de la population et dont

la variance est égale à la variance de la population divisée par le nombre d'échantillons $\left(\mu_s = \mu ; \sigma_s^2 = \sigma^2 / n_s\right)$. De plus, la moyenne de plusieurs observations aléatoires tend à suivre une distribution normale $N\left(\mu_S, \sigma_S\right)$ lorsque le nombre d'observations est suffisamment grand, même si les observations effectuées ne suivent pas une distribution normale [87].

Sur base de ces deux affirmations, PAGE utilise le *fold change* comme statistique individuelle. Le *fold change* n'est pas distribué suivant une normale, mais lorsque les *fold change* associés à tous les gènes d'un groupe sont utilisés pour calculer un *fold change* moyen, celui-ci est distribué suivant une normale. Le théorème central limite permet en outre de tenir compte de la taille du groupe de gènes lors du calcul du score Z (Equ. II.B.19 et II.B.20). La *p-value* associée au score Z est ensuite attribuée sur base d'une distribution normale [87].

$$FC_{mean} = \frac{\sum_{i=0}^{n*G} FC_i}{n_{*G}} \sim N\left(\mu_{*G} = \mu_{**}, \sigma_{*G} = \frac{\sigma_{**}}{n_{*G}}\right) \quad \text{(Equ. II.B.19)}$$

$$Z = \frac{\left(FC_{mean} - \mu_{**}\right) n_{*G}}{\sigma_{**}} \sim N(0,1) \quad \text{(Equ. II.B.20)}$$

Les auteurs ont montré que la distribution du *fold change* moyen suit une distribution normale lorsqu'au moins dix gènes sont membres du groupe étudié. Le principal avantage de la méthode provient de la facilité avec laquelle la significativité est attribuée, nécessitant un temps de calcul négligeable en comparaison des méthodes non paramétriques qui utilisent des permutations [87].

Bien que cela n'ait pas été réalisé par les auteurs, la méthodologie suivie peut être appliquée pour d'autres statistiques individuelles (par exemple la statistique *t* de STUDENT, ou l'une de ses variantes), au lieu du *fold change*.

II.B.4.c. Les statistiques absmean et maxmean

EFRON & TIBSHIRANI ont introduit, en 2007, l'utilisation de la statistique *maxmean*. Au départ de scores individuels tels que le *t* de STUDENT, ils tiennent compte de la direction de la

statistique individuelle pour définir deux ensembles de valeurs $\left(s^{(+)}; s^{(-)}\right)$. Ces deux ensembles sont ensuite utilisés pour calculer deux valeurs moyennes distinctes pour les gènes respectivement sur-exprimés et sous-exprimés. La statistique la plus élevée entre ces deux valeurs est utilisée pour caractériser le groupe de gènes (Equations II.B.21 et II.B.22) [53].

$$\bar{s}^{(+)}_{*G} = \frac{\sum_{i=0}^{n*G} s_i^{(+)}}{n_{*G}} \quad \text{et} \quad \bar{s}^{(-)}_{*G} = \frac{\sum_{i=0}^{n*G} s_i^{(-)}}{n_{*G}}$$

$$\text{avec } s_i^{(+)} = max(t_i, 0) \text{ et } s_i^{(-)} = -min(t_i, 0) \quad \text{(Equ. II.B.21)}$$

$$S_{maxmean} = max\left\{\bar{s}^{(+)}_{*G}, \bar{s}^{(-)}_{*G}\right\} \quad \text{(Equ. II.B.22)}$$

La méthode permet donc d'étudier des groupes de gènes comprenant simultanément des gènes sur- et sous-exprimés. Cependant, dans pareil cas, seul les gènes appartenant au sous-groupe dont la moyenne est la plus élevée sont considérés. Les auteurs décrivent également la statistique *absmean*, dont le principe est similaire, mais repose sur le calcul de la moyenne de la valeur absolue de la statistique *t* dans chacun des sous-groupes directionnels, pour en choisir la valeur maximale [53]. Celle-ci- ne tient donc pas compte de la taille relative des deux sous-groupes, contrairement à la statistique *maxmean*.

La significativité du test est ensuite attribuée, dans les deux cas, sur base d'une procédure en deux étapes appelée *restandardization*. La première étape vise à corriger la statistique calculée, sur base de permutations de gènes. La seconde étape évalue la significativité sur base de permutations d'échantillons. Cette procédure a été mise au point par analogie avec l'approche proposée par Subramanian et al. (GSEA [131]) quelle que soit la statistique utilisée pour caractériser le groupe de gènes, et est donc applicable aux autres méthodes. Dans le cas de modèles simples tels que l'utilisation d'une statistique moyenne, la procédure se simplifie car la première étape peut y être calculée sur base des estimateurs individuels de la moyenne et de la variance [53].

II.B.4.d. SAMGS et la somme quadratique de la statistique d

Dinu *et al.*, en 2007 [44], se basent sur les travaux réalisés par Tusher *et al.* en analyse

individuelle (méthode SAM [136]) et sur les méthodes multivariées basées sur le T^2 de HOTTELING [70], et l'approche décrite par DEMPSTER (1958, 1960) [40, 41]. La statistique individuelle utilisée correspond à la statistique d, variante du t du STUDENT impliquant un terme correctif de la variance (*fudge factor*) (Equation II.A.30 p. 52). Les auteurs tiennent compte du modèle de DEMPSTER, et de la définition de la statistique WD (*Weighted DEMPSTER*) pour caractériser un groupe de gènes (Equation II.B.23) [44].

$$WD = \frac{\sum_{i=1}^{|n*G|} d_i^2}{\hat{E}\left[\sum_{i=1}^{|n*G|} d_i^2\right]} \quad \text{(Equ. II.B.23)}$$

Le dénominateur de l'équation II.B.23 représente la moyenne de n_1+n_2-2 valeurs de même moyenne et variance que le numérateur, obtenues par une transformation orthonormale des valeurs d'expression associées au groupe de gènes. Ce dénominateur peut être omis dans l'expression finale de la statistique $SAMGS$, car la significativité y est calculée sur base de permutations d'échantillons, fournissant la même valeur pour $SAMGS$ et pour WD car le dénominateur de WD est constant, si bien que la statistique employée correspond à la somme du carré de la statistique d individuelle (Equation II.B.24) [44].

$$SAMGS = \sum_{i=1}^{|n*G|} d_i^2 \quad \text{(Equ. II.B.24)}$$

Les auteurs justifient l'utilisation de la somme des d^2, par comparaison à la somme de la statistique d, car la méthode s'avère plus apte à détecter des groupes de gènes au sein desquels moins de 30% des gènes présentent une différence d'expression (sur base de simulations) [44]. La correction pour tests multiples est réalisée en suivant l'approche développée par STOREY, qui met en oeuvre le calcul d'une *q-value* [128].

L'un des avantages de la méthode, comparée aux méthodes basées sur un score d'enrichissement, est que l'analyse d'un groupe de gènes ne nécessite que les données relatives à ce groupe. Cette affirmation doit toutefois être relativisée puisque l'ensemble du jeu de données est utilisé pour calculer le terme de correction de la variance de la statistique individuelle. Néanmoins, les difficultés d'interprétation soulevées par les méthodes hybrides n'affectent pas SAMGS. De plus, l'utilisation d'une somme quadratique assure le cumul des réponses individuelles, indépendamment de leur direction [44].

II.B.5. Généralisation de la stratégie post-hoc, et améliorations

II.B.5.a. Introduction

BARRY *ET AL.* présentent une généralisation des travaux de VIRTANEVA (qui utilisent la statistique de WILCOXON [140]) et de MOOTHA (Score d'enrichissement [106]) [13]. La stratégie générale de l'analyse repose sur la combinaison de deux statistiques arbitrairement choisies pour l'analyse individuelle (U, par exemple le t de STUDENT) et l'analyse de groupes de gènes (V, par exemple la statistique de KOLMOGOROV-SMIRNOV ou celle de WILCOXON) [13]. Cependant, les deux stratégies *post-hoc* reposent sur des postulats différents, compétitifs ou auto-suffisants. Il est donc important de tenir compte des hypothèses testées par les différentes méthodes, particulièrement en regard du mode de permutations utilisées, et de l'étape à laquelle elles interviennent.

II.B.5.b. Hypothèses et permutations

Tester l'association entre un groupe de gène et le phénotype étudié correspond à deux hypothèses :

- **Q1**: Les gènes membres du groupe étudié ne présentent pas une association significative avec le phénotype, en comparaison avec les autres gènes.
- **Q2**: Aucun gène du groupe ne présente une association entre son expression et le phénotype étudié.

La différence fondamentale entre ces deux hypothèses repose sur la manière d'étudier l'association entre les gènes et le phénotype. L'hypothèse Q1 compare l'association expression/phénotype pour un groupe de gènes avec cette association pour les autres gènes, tandis que l'hypothèse Q2 compare uniquement le phénotype avec l'expression des gènes au sein du groupe.

TIAN *ET AL.*, en 2005, proposent d'aborder séparément ces deux questions, sur base des statistiques T et E, respectivement pour l'hypothèse compétitive (Q1) et l'hypothèse auto-suffisante (Q2)(Equation II.B.25). La statistique T est ensuite évaluée sur base d'une procédure de permutations de gènes (distribution de T_{perm}), tandis que la statistique E

est évaluée sur base d'une procédure de permutations d'échantillons (distribution E_{perm}) [132].

$$T_G = E_G = \frac{1}{n_{*G}} \sum_{i=1}^{n*G} t_{i \in G} \quad \text{(Equ. II.B.25).}$$

La même valeur statistique ($T = E$) est donc évaluée d'une part en fixant le phénotype et en comparant des groupes aléatoires, d'autre part en fixant la définition du groupe et en comparant des échantillons aléatoires, en accord avec les questions Q1 et Q2 respectivement. Ils proposent aussi l'utilisation de poids différents pour chaque gène lors du calcul de la statistique T (ou E), et permettre ainsi de tenir compte des corrélations entre les valeurs *t* individuelles (Equ. II.B.26) [132].

$$T_G = E_G = \frac{1}{n_{*G}} \sum_{i=1}^{n_{*G}} w_i t_{i \in G} \quad \text{(Equ. II.B.26).}$$

Puisque chaque groupe de gène est caractérisé par une structure différente (nombre de gènes, corrélations...), la distribution des statistiques T et E pour chaque groupe est différente, et une étape de standardisation est nécessaire pour pouvoir comparer les différents groupes de manière objective [132].

II.B.5.c. La procédure de « restandardization »

EFRON & TIBSHIRANI, en 2007, mettent en évidence l'inadéquation de la procédure de permutations utilisée dans GSEA et les méthodes dérivées [53]. Ils nomment *randomization* la procédure basée sur des permutations de gènes, et *permutations* la procédure basée sur des permutations d'échantillons. Partant d'un jeu de données simulées, ils montrent que lorsque, pour chaque groupe de gènes, la moitié des gènes sont différentiellement exprimés, avec la même magnitude, tous les scores paraissent significatifs en comparaison de la distribution obtenue par permutations de gènes, mais aucun n'apparaît différent de groupes générés aléatoirement (en contradiction avec l'hypothèse compétitive). Dès lors, ils proposent d'adapter la procédure généralisée (quelles que soient les statistiques individuelles et de groupe), pour tenir compte simultanément des deux schémas de permutations. La procédure de *restandardization* consiste à standardiser dans une première étape les valeurs de la statistique de groupe grâce aux valeurs obtenues par permutations de gènes, et d'ensuite calculer la significativité sur base

d'une distribution générée par permutations d'échantillons [53].

Selon l'hypothèse nulle $H_{0,perm}$, les échantillons présentent n_{*G} valeurs indépendantes qui sont distribuées identiquement (i.i.d.). En revanche, l'hypothèse nulle associée à la permutation de gènes, $H_{0,rand}$, postule que le groupe de gènes est constitué par une sélection aléatoire de n_{*G} gènes au départ du jeu complet. Le biais observé provient de l'existence de corrélations entre les différents gènes, inhérentes aux propriétés biologiques de l'échantillon. Les permutations d'échantillons éliminent la corrélation entre les gènes. Cette procédure fournit un score de groupe pour lequel la variabilité est sous-estimée lorsqu'il existe une corrélation entre ses membres [53].

La formulation générale de la procédure de *restandardization* est illustrée par l'équation II.B.27.

$$V^{**} = \mu^r + \frac{\sigma^r}{\sigma^p}\left(\frac{V^p - \mu^p}{\sqrt{n_{*G}}}\right) \quad \text{(Equ. II.B.27)}$$

avec μ^r et σ^r, la moyenne et déviation standard de la distribution de la statistique de groupe V^r obtenue par *randomization* (permutations de gènes), μ^p et σ^p, la moyenne et déviation standard de la distribution de la statistique de groupe V^p obtenue par permutations d'échantillons [53].

Les auteurs envisagent ensuite des cas de figures où la définition de la statistique de groupe permet de simplifier le modèle. En pratique, cette procédure, implémentée dans le logiciel GSA (Gene Set Analysis), repose sur 4 étapes :

- le calcul d'une statistique individuelle (U) ;
- le calcul d'une statistique de groupe (V) ;
- la standardisation de la statistique de groupe $V^r = \left(V - \mu^r\right)/\sigma^r$ (permutations de gènes) ;
- l'application des 3 premières étapes sur des permutations d'échantillons.

Les valeurs obtenues par permutations permettent alors d'estimer la significativité des scores obtenus [53].

II.B.6. Les méthodes « globales »

II.B.6.a. Introduction

Une troisième catégorie de méthodes, appelées méthodes globales, analysent les données d'expression propres à un groupe de gènes dans leur globalité, sans passer par une étape intermédiaire d'analyse individuelle. Les modèles utilisés par ces méthodes sont multivariés, et la structure particulière des données propres aux groupes y est utilisée.

II.B.6.b. GlobalTest

Sur base d'un modèle bayesien empirique, GOEMAN *ET AL.*, en 2003, ont développé la méthode *GlobalTest*, appliquant un modèle linéaire généralisé à la problématique de l'analyse de groupes de gènes [63].

Le postulat de départ est que si un groupe de gènes peut être utilisé pour prédire l'issue clinique des échantillons étudiés, alors les données d'expression doivent être différentes pour différents cas cliniques. La modélisation de la dépendance entre le phénotype et les données d'expression du groupe peut être exprimée sur base d'un modèle linéaire généralisé. L'équation II.B.28 présente la formulation de tels tests, ou α représente l'ordonnée à l'origine, et β est le vecteur des coefficients de régression pour un gène donné. La fonction h est une fonction de liaison [63].

$$E\left(Y_j|\beta\right)=h^{-1}\left(\alpha+r_j\right)=h^{-1}\left(\alpha+\sum_{i=1}^{n_{*G}} x_{ij}\beta_i\right) \text{ (Equ. II.B.28).}$$

avec Y, le vecteur de description phénotypique et x_{ij} les valeurs d'expressions individuelles au sein du groupe de gènes. Dans ce contexte, l'hypothèse nulle est que le groupe de gène n'a pas de pouvoir prédictif, et consiste à vérifier que tous les coefficients de régression sont nuls (Equation II.B.29) [63].

$$H_0: \beta_1=\beta_2=...=\beta_{n_{*G}}=0 \text{ (Equ. II.B.29)}$$

Cette hypothèse nulle peut être testée si l'on considère que les coefficients de régression β

sont issus d'un distribution commune $(0,\tau^2)$, où τ^2 est la variance caractérisant la déviation entre les coefficients β et 0. Dès lors, l'hypothèse nulle est reformulée par l'équation II.B.30 [63].

$$H_0: \tau^2 = 0 \quad \text{(Equ. II.B.30)}$$

La statistique employée pour tester l'hypothèse peut être formulée de façon équivalente par trois équations, permettant une interprétation plus simple des résultats (Equations II.B.31 à II.B.33) [63].

$$Q = \frac{(Y-\mu)' R (Y-\mu)}{\mu_2} \quad \text{(Equ. II.B.31)}$$

où $\mu = h^{-1}(\alpha)$ est la valeur attendue de Y et μ_2 est le second moment central de Y lorsque l'hypothèse nulle est vérifiée, $R = (1/n_{*G}) X X'$ est une matrice de taille $n \times n$ proportionnelle à la matrice de covariance des effets aléatoires r (matrice de covariance entre les échantillons). X est la matrice des valeurs d'expression du groupe de gènes [63].

En reformulant l'équation II.B.31, nous pouvons constater que la statistique Q caractéristique d'un groupe de gènes est la moyenne arithmétique des statistiques Q_i individuelles calculées pour chaque gène du groupe. Chaque statistique individuelle Q_i est un multiple du carré de la covariance entre le pattern d'expression du gène i et le phénotype (résultat clinique) (Equ. II.B.32). Les gènes pour lesquels la variance est la plus élevée influencent donc plus fortement la statistique Q caractéristique du groupe de gènes.

$$Q = \frac{1}{n_{*G}} \sum_{i=1}^{n*G} Q_i = \frac{1}{n_{*G}} \sum_{i=1}^{n*G} \frac{1}{\mu_2} \left[X_i' (Y-\mu) \right]^2 \quad \text{(Equ. II.B.32)}$$

Enfin, en reformulant l'expression de la statistique Q sur base de l'équation II.B.33, on constate qu'elle se calcule par la somme sur tous les termes du produit de deux matrices de covariance : D'une part R_{ij} la matrice de covariance des patterns d'expressions entre les gènes, et d'autre part $(Y_i-\mu)(Y_j-\mu)$ la matrice de covariance du phénotype des échantillons (résultat clinique). Autrement dit, la valeur Q est élevée si les structures de covariance entre les gènes et entre les échantillons sont similaires. La procédure *GlobalTest* vérifie donc si les échantillons dont l'expression des gènes est similaire présentent

également un même phénotype [63].

$$Q=\frac{1}{\mu_2}\sum_{i=1}^{n}\sum_{j=1}^{n}R_{ij}(Y_i-\mu)(Y_j-\mu) \quad \text{(Equ. II.B.33)}$$

Les auteurs montrent que la statistique Q est distribuée asymptotiquement suivant une distribution normale. Elle correspond également à une statistique quadratique, et peut par conséquent être évaluée sur base d'une distribution $c\chi^2_\nu$, où c est un facteur de mise à l'échelle et ν est le nombre de degrés de liberté. Cette approche fournit une meilleure approximation sur des échantillons de petite taille. Enfin, la significativité peut être attribuée sur base de permutations d'échantillons [63].

II.B.6.c. GlobalAncova

La méthode *GlobalTest*, décrite ci-dessus, utilise un modèle linéaire généralisé pour tester la dépendance du phénotype Y vis-à-vis des mesures d'expression individuelles des gènes X. L'hypothèse testée est donc formulée par l'équation II.B.34 [63].

$$P(Y|X)=P(Y) \quad \text{(Equ. II.B.34)}$$

Mansmann & Meister, en 2005, critiquent cette approche, car dans le contexte de l'analyse de l'expression différentielle, la question qui est posée est formulée à l'inverse : « les mesures d'expressions observées sont-elles dues aux phénotypes comparés ? », formulée par l'équation II.B.35 [104].

$$P(X|Y=1)=P(X|Y=2) \quad \text{(Equ. II.B.35)}$$

En cas de respect de l'hypothèse nulle, le théorème de Bayes montre que les deux hypothèses sont équivalentes (Equation II.B.36).

$$\frac{P(X|Y=1)}{P(X|Y=2)}=\frac{P(Y=1|X)}{P(Y=2|X)}\times\frac{P(Y=2)}{P(Y=1)}=\frac{P(Y=1)}{P(Y=2)}\times\frac{P(Y=2)}{P(Y=1)}=1 \quad \text{(Equ. II.B.36)}$$

Mansmann & Meister, dans la méthode *GlobalAncova*, formulent l'hypothèse nulle par l'intersection des hypothèses nulles associées à chacun des gènes (Equation II.B.37). Ils utilisent un modèle *ANCOVA* pour étudier l'effet de l'interaction ou l'effet de l'interaction et de l'effet principal du groupe. Le modèle tient compte d'un cofacteur, attribué pour chaque gène, échantillon par échantillon. Celui-ci indique si les mesures du gène doivent être

corrigée spécifiquement. A titre d'exemple, si le critère choisi est le sexe, tous les gènes liés au sexe seront corrigés en tenant compte uniquement de la moyenne propre aux mesures réalisées sur les individus du même sexe. L'équation II.B.38 présente le modèle complet utilisé par *GlobalAncova* [104].

$$H_{0\mathrm{i}}: \mu_{1,i} = \mu_{2,i} \quad \text{(Equ. II.B.37)}$$

où i est l'indice du gène.

$$\begin{aligned} X_{ijk} &= \mu_{ij} + \theta_j z_{ik} + \varepsilon_{ijk}\,;\, E\left(\varepsilon_{ijk}\right) = 0 \\ \mu_{ij} &= \mu + \alpha_i + \beta_j + \gamma_{ij} \end{aligned} \quad \text{(Equ. II.B.38)}$$

où z_{ik} symbolise le cofacteur associé au gène i dans l'échantillon k, α_i, β_j et γ_{ij} symbolisent les effets du groupe, de la condition, et de leur interaction, respectivement. ε_{ijk} symbolise l'erreur résiduelle. Lorsque l'hypothèse nulle est vérifiée, le modèle se simplifie car $\alpha_i = \beta_j = \gamma_{ij} = 0$, et le modèle, qualifié de « réduit », s'exprime par $\mu_{ij} = \mu$ [104].

L'hypothèse nulle étant stricte (aucun gène ne présente une moyenne différente entre les deux conditions), l'interaction est par conséquent nulle lorsque l'hypothèse nulle est respectée. Il est donc possible de tester l'hypothèse nulle sur base de l'interaction, comme le présente l'équation II.B.39, qualifiée de « modèle additif » [104].

$$\begin{aligned} H_0': \gamma_{ij} &= 0 \\ \mu_{ij} &= \mu + \alpha_i + \beta_j \end{aligned} \quad \text{(Equ. II.B.39)}$$

Les hypothèses nulles définies sont testées en accord avec les procédures *ANOVA* habituelles, en calculant la statistique *F* associée, sur base des équations II.B.40 et II.B.41.

$$H_0: F = \frac{\left(\frac{SSR_{Reduced} - SSR_{Full}}{df_1}\right)}{\left(\frac{SSR_{Full}}{df_2}\right)} \quad \text{(Equ. II.B.40)}$$

$$H_0': F = \frac{\left(\frac{SSR_{Add} - SSR_{Full}}{df_1'}\right)}{\left(\frac{SSR_{Full}}{df_2}\right)} \quad \text{(Equ. II.B.41)}$$

où $df_1 = n, df_1' = \left(df_1 - 1\right), df_2 = n\left(m_1 + m_2 - \left(p+2\right)\right), n$ est le nombre de gènes, m_1 et m_2 sont les nombres de mesures dans les deux conditions, et p est le nombre de cofacteurs.

Enfin, la significativité du test n'est pas estimée en utilisant la distribution F traditionnelle, car la corrélation possible entre les membres du groupe de gènes et la probable non homogénéité des variances conduit à une sous-estimation de la *p-value*. La procédure d'évaluation de la significativité repose dès lors sur la permutation des échantillons [104].

GlobalAncova est donc une méthode similaire à l'*ANOVA*, et le test effectué sur H_0 correspond à l'hypothèse stricte, mais permet de corriger les données sur base d'un cofacteur. MANSMANN & MEISTER pensent toutefois que l'hypothèse H_0' est plus intéressante, car elle correspond à une réalité biologique, qui n'est ni globale, ni unidirectionnelle, ni homogène [104].

> ✍ Bien que la critique de l'hypothèse nulle de *GlobalTest* soit justifiée, l'examen de l'équation II.B.32 montre que la statistique Q de *GlobalTest* est la moyenne des statistique Q individuelles, chacune étant relative à la réponse du gène considéré. Mathématiquement, la critique n'est donc pas justifiée.

III. Objectifs

Les recherches que nous avons menées depuis l'initiation du projet reposent sur une démarche multiple et permanente.

Déminer

Le premier objectif vise à dresser une liste comparative des méthodes disponibles, et à la tenir à jour sur base des études récentes, afin de dégager des similitudes et enseignements généraux, pour déminer le terrain de l'analyse de l'expression différentielle, et identifier *a priori* les démarches adéquates qui améliorent les performances des tests réalisés.

Proposer

Sur base de la compréhension des limites des méthodes disponibles, principalement liées au nombre de mesures réalisées, le second objectif consiste à proposer une amélioration des procédures classiques. En terme d'analyse individuelle, nous proposons d'estimer la variance individuelle sur base du partage d'information entre les gènes, pour augmenter le nombre de mesures considérées et repousser ainsi les limites des procédures statistiques classiques. Nous proposons également l'ajout d'une étape additionnelle d'analyse globale, qui repose sur la combinaison des résultats de plusieurs méthodes pour obtenir des résultats plus fiables. En terme d'analyse de groupe de gènes, nous proposons l'utilisation d'une procédure multivariée du type *ANOVA-2*, ainsi qu'une procédure bidirectionnelle dérivée, pour améliorer les performances de l'analyse.

Tester

Pour adapter au mieux les démarches proposées, le troisième objectif visé est l'étude du comportement des statistiques proposées, en regard de l'erreur commise sur l'estimation des paramètres, du nombre de mesures, de la définition des estimateurs, et d'identifier ainsi les conditions optimales d'usage des estimateurs proposés, pour définir de nouvelles méthodes d'analyse.

Valider

Par comparaison *a posteriori* aux démarches proposées dans d'autres études, le quatrième objectif poursuivi vise à quantifier les avantages et inconvénients, en terme de performances, associés à chacune des démarches publiées, et de valider notre démarche en quantifiant les améliorations proposées.

Automatiser

L'utilisation et la paramétrisation des analyses de l'expression différentielle nécessitant une compréhension de plusieurs méthodes statistiques, et des différentes étapes de l'analyse, nous avons pour objectif d'automatiser les enseignements de nos recherches au sein d'un outil logiciel simple. Pour les utilisateurs non statisticiens, cet outil vise à automatiser les choix nécessaires pour optimiser les analyses, et à guider l'utilisateur. Pour les bioinformaticiens et biostatisticiens, cet outil permet une paramétrisation manuelle et une adaptation sur mesure de la stratégie analytique, en combinant diverses démarches réalisées à des étapes différentes, et permet l'évaluation des performances de la démarche envisagée.

IV. Résultats

IV.A.
Analyse de l'expression différentielle par gène

Résumé

Ce chapitre présente les recherches que nous avons menées sur la thématique de l'analyse des données d'expression relatives aux gènes.

En initiant ce projet, en 2004, les méthodes d'analyse principalement utilisées dans la littérature reposaient sur le *fold change*, le test de STUDENT, *SAM* (TUSHER *ET AL.*, 2001 [136]), et le *regularized t-test (*BALDI *ET AL.*, 2001 [11]*)*.

Le postulat à l'origine du projet repose sur l'amélioration de l'estimation de la variance, grâce à un partage d'information entre les gènes représentés au sein du jeu de données analysé, et l'utilisation de critères valides pour guider les analyses réalisées.

D'un point de vue méthodologique, ce postulat correspond à la définition d'un ensemble de gènes (la fenêtre), qui seront utilisés conjointement lors de l'estimation de la variance individuelle, utilisée en lieu et place de l'estimateur classique.

Afin de tester ces idées, nous débutons la présentation de nos résultats par l'étude du critère utilisé au sein du *regularized t-test* : une relation empirique qui relie la variabilité individuelle et le niveau d'expression. En utilisant ce critère connu comme point de départ pour valider l'utilisation d'une fenêtre, nous avons pu dégager plusieurs enseignements sur l'effet bénéfique de l'estimateur proposé. Nous décrivons ensuite la validation de la méthode, dénommée *window t-test*, par comparaison de ses performances avec les meilleurs méthodes actuellement disponibles. Les évaluations ont été réalisées sur trois types de jeux de données (simulations, *spike-in* et biologique) et illustreront que la méthode *window* atteint le niveau de performance des meilleures méthodes actuelles, et le dépasse pour plusieurs des conditions testées. A titre d'exemple, une expérience prévue pour cinq réplicats peut être réalisée sur deux réplicats avec le même niveau de performances.

Nous clôturons ensuite cet exposé des résultats de l'analyse individuelle en proposant l'ajout d'une nouvelle étape dans la méthodologie : l'évaluation d'un *consensus* sur plusieurs méthodes. Nous avons testé cette idée sur base d'une formulation mathématique simple, pour offrir la possibilité de combiner les résultats des meilleures méthodes et d'améliorer ainsi les performances. Choisir une méthode d'analyse est souvent un choix gouverné par la structure des données, chaque méthode ayant ses faiblesses. Les résultats montrent, en fin de ce chapitre, que le *consensus*, évalué en présence de méthodes performantes, de méthodes traditionnelles, et de méthodes moins performantes, permet d'obtenir des résultats fiables.

IV.A.1. Introduction

L'analyse de l'expression différentielle individuelle pose le défi de réaliser un grand nombre de tests d'hypothèse sur des jeux de données caractérisés par un nombre restreint de mesures. L'estimation de la variance sur un petit nombre de mesures est peu fiable. L'amélioration la plus immédiate pour obtenir des résultats plus fiables repose donc *a priori* sur l'usage d'un plus grand nombre de données.

Pour compenser les limites posées par le nombre de tests réalisés avec peu de mesures, l'objectif global de cette thèse repose sur l'utilisation de plusieurs gènes, liés biologiquement, pour augmenter le nombre de mesures utilisées lors de l'estimation de la variance individuelle. Si les critères choisis pour « regrouper » les gènes sont valides, l'utilisation de données supplémentaires associées à ces gènes est associée à une sensibilité et à une spécificité plus importantes. Nous pensons ainsi repousser les limites statistiques posées par la structure des données, pour tirer parti des liens existant entre les gènes, c'est à dire tenir compte de la structure biologique des données.

Pour formuler statistiquement cette idée, la méthodologie que nous allons présenter repose sur l'utilisation d'une procédure statistique classique, le test du t de STUDENT (procédure univariée), utilisée avec des estimateurs plus fiables de la variance, calculé sur plusieurs gènes. Nous appelons « fenêtre » le groupe de gènes utilisé pour extraire des informations relatives à la variabilité, en complément du gène étudié.

Plusieurs critères peuvent être utilisés pour définir le groupe de gènes, choisir les gènes les plus à même de partager une variabilité similaire à celle associée au gène d'intérêt. Parmi les publications disponibles aux premiers jours des recherches menées, deux d'entre elles, la méthode ***LPE*** et la méthode *regularized t-test*, reposent sur l'utilisation d'une relation empirique entre le niveau d'expression et la variabilité.

A terme, l'utilisation de la fenêtre sur base de la définition de groupes de gènes reliés par leur participation à une même voie métabolique, ou régulés par un même facteur, pourraient faciliter la démarche. Cependant, la complexité des relations biologiques, et les connaissances biologiques incomplètes, ne permettent pas encore de prédire de manière exhaustive les corrélations entre les gènes. Cette information est obtenue empiriquement, sur base des expériences réalisées en laboratoire. N'ayant pas connaissance du lien entre la structure biologique et la structure mathématique des données, la relation entre le niveau d'expression moyen et la variabilité nous semble un bon point de départ pour investiguer

les capacités d'une telle méthode, car il s'agit d'une observation empirique d'un critère utilisable rapporté dans la littérature scientifique. Nous montrerons que l'utilisation d'une fenêtre est une méthode efficace lorsque le nombre de mesures disponibles est limité (moins de 5), que les performances de la méthode surpassent ou égalent les performances des autres méthodes disponibles. La suite logique consisterait à appliquer ensuite la même approche sur d'autres critères, mais une compréhension de la structure des groupes connus est nécessaire pour les définir. La seconde partie de nos recherches entame l'étude des groupes, afin que, dans les recherches qui seront menées à l'avenir, les outils et la démarche développée en analyse de groupes puissent identifier un ou plusieurs critères de choix à utiliser en combinaison avec la fenêtre pour améliorer l'analyse individuelle.

IV.A.2. La méthode « window t-test »

IV.A.2.a. Etude de la relation entre la variabilité et le niveau d'expression

Afin d'illustrer la relation empirique entre le niveau d'expression et la variabilité individuelle, et d'initier l'étude de la qualité de l'estimation de la variance individuelle sur base de plusieurs gènes, nous avons choisi d'utiliser le jeu de données E-MEXP-231, qui contient 49 réplicats de puces à ADN hybridées avec des échantillons d'adenocarcinomes primaires du poumon.

Le choix de ce jeu repose sur le nombre de mesures disponibles, qui permet donc d'un point de vue statistique de calculer des estimateurs fiables de la moyenne et de la variance. Des intervalles réguliers de niveau d'expression ont été définis sur base du niveau moyen d'expression associé à chaque gène. Pour chaque intervalle, nous avons ensuite calculé la valeur médiane de la variance individuelle, ainsi que ses quartiles (seuils en dessous desquels se distribuent 25% et 75% de la distribution des variances dans l'intervalle). La valeur médiane est utilisée pour illustrer la relation avec le niveau d'expression. Les estimateurs individuels, représenté en gris, et les quartiles, permettent de visualiser le niveau de dispersion des variances individuelles autour de cette relation empirique. L'ensemble de ces indicateurs est représenté dans la figure IV.A.1, pour les deux méthodes de prétraitement les plus couramment utilisées (MAS 5.0 et GCRMA).

La première observation qu'il convient de rapporter repose sur la dispersion des points observée par rapport à une tendance centrale liée au niveau d'expression. Nous pouvons conceptualiser la variance sur base de deux termes, relatifs respectivement à la variabilité individuelle et à la variabilité due au niveau d'expression, pour exprimer la dispersion des points observée dans la figure IV.A.1, qui est centrée autour d'une valeur qui évolue avec le niveau d'expression.

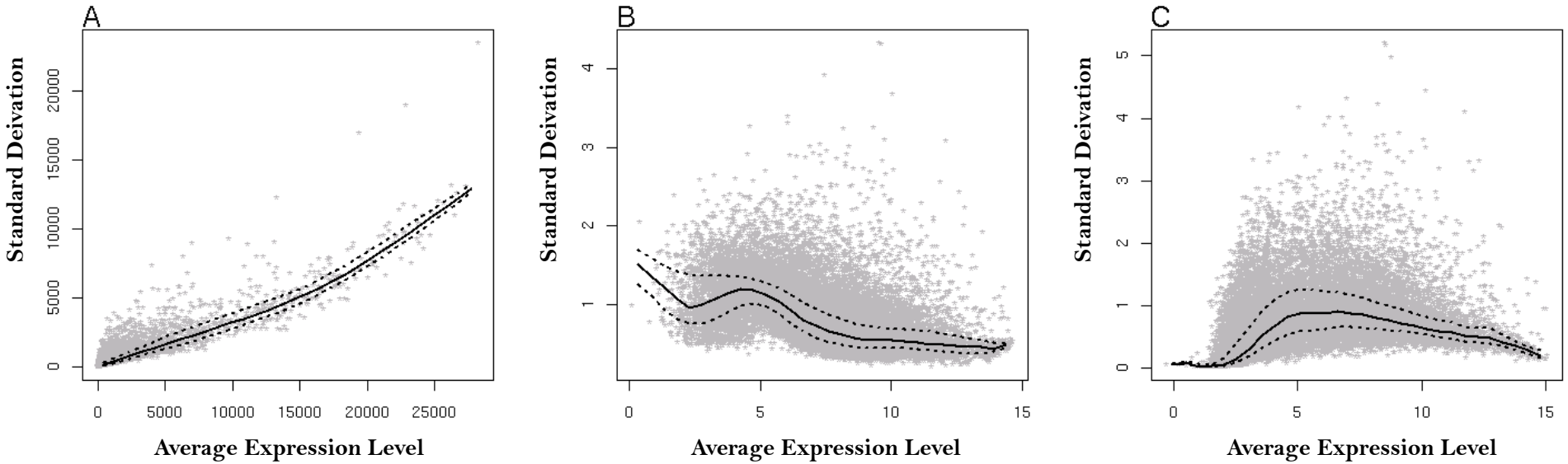


Figure IV.A.1 : Illustration de la relation empirique entre le niveau d'expression moyen et la variabilité. L'écart-type et la moyenne des données d'expression ont été calculés pour chaque *probeset* du jeux de données E-MEXP-231 (en gris, 49 réplicats relatifs à l'adénocarcinome primaire du poumon), pour différents prétraitements. A : MAS 5.0 ; B: MAS 5.0 (Log 2) ; C: GCRMA. Les mesures obtenues ont été utilisées pour définir 30 intervalles réguliers de niveau d'expression. Les valeurs médianes (traits pleins) et les quartiles de la distribution de l'écart-type (traits discontinus) ont été calculés pour chaque intervalle. La ligne en train plein illustre donc la relation empirique entre le niveau d'expression et la variabilité individuelle. Les lignes en traits discontinues offrent un aperçu de la dispersion de la variabilité individuelle autour de cette tendance centrale.

La comparaison de ces indicateurs entre les deux prétraitement envisagés montre que cette relation entre la variabilité et le niveau d'expression est affectée, pour un même jeu de données, par la préparation des données réalisée au cours du prétraitement. Nous pouvons entre autres y observer que l'utilisation de GCRMA conduit à une variabilité proche de 0 pour un niveau d'expression nul, contrairement à MAS 5.0. La comparaison de la même relation avec d'autres méthodes présente un intérêt en tant que tel, mais nous limiterons nos recherches à ces deux méthodes, car d'une part elles sont très répandues, et d'autre part notre objectif méthodologique concerne l'intégration de données supplémentaires, et non l'analyse du choix optimal de la méthode de prétraitement, qui constitue un sujet d'étude en soit. L'utilisation de GCRMA sera toutefois surveillée tout au long de la première partie, par comparaison avec MAS 5.0, car la méthode repose sur l'utilisation d'un critère biologique connu : la proportion de nucléotides G et C dans la séquence génétique, susceptible de s'hybrider plus fortement que les nucléotides A et T.

L'observation de la figure IV.A.1, sur un même jeu de données, entre les méthodes de prétraitement envisagées, suggère que la prédiction d'une relation mathématique simple pour prédire la variance associée à un gène n'est pas envisageable. La première raison en est qu'elle diffère entre les méthodes. La seconde raison en est la dispersion importante des points autour de cette relation. N'ayant pas de modèle complet pour expliquer la variabilité, deux hypothèses extrêmes peuvent être émises :

- la dispersion des points autour de la relation est due au hasard (H_0) ;
- la dispersion des points autour de la relation est due aux propriétés individuelles des gènes (H_1).

Si l'hypothèse H_0 est vérifiée, cela signifie que la variabilité peut tout simplement être estimée sur base d'une variabilité moyenne associée au niveau d'expression. Dans le cas contraire, cela signifie que la variabilité individuelle diffère de la relation moyenne-variance, et qu'une procédure adaptée ne pourrait que partiellement exploiter la relation observée pour « stabiliser » la variance, en évitant des valeurs trop éloignées de la relation suivie globalement par les données.

Statistiquement, la relation empirique observée peut être formulée par l'équation IV.A.1, qui défini un modèle tenant compte de l'effet du niveau d'expression sur la variabilité.

$$X = \mu + a + \varepsilon \quad \text{(Equ. IV.A.1)}$$

où a symbolise l'effet du niveau d'expression et ε symbolise l'erreur résiduelle.

Puisqu'une relation mathématique directe de prédiction de l'effet du niveau d'expression, symbolisé par a, n'est pas disponible, nous pensons que l'utilisation d'une fenêtre permet de capturer une partie de cet effet au cours de l'estimation de la variance. En utilisant un grand nombre de gènes qui partagent le même niveau d'expression pour définir cette fenêtre, la variabilité s'approche de plus en plus de la variabilité due au niveau d'expression (le cas extrême correspond à l'ensemble du jeu de données, et à une variance unique pour tous les gènes). Le prochain paragraphe présente donc l'utilisation de la fenêtre pour estimer cette variabilité, avant de l'utiliser pour caractériser la qualité et le bien fondé de l'utilisation de ce critère pour estimer la variance.

IV.A.2.b. Calcul de la variance au départ d'une fenêtre définie par le niveau d'expression

La figure IV.A.2 illustre la procédure utilisée pour estimer la variance au départ de plusieurs gènes situés à un même niveau d'expression.

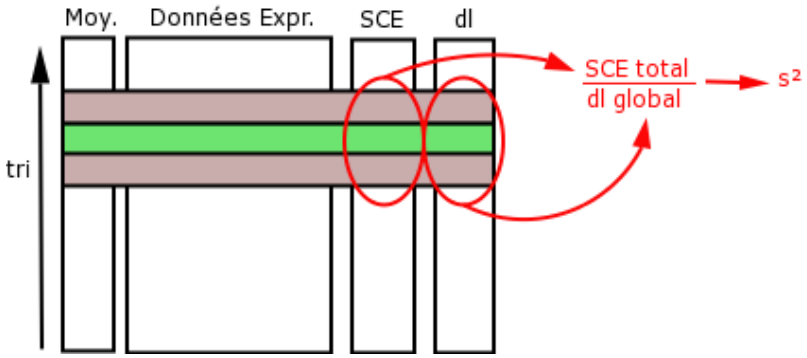


Figure IV.A.2 : Illustration de la procédure de calcul des estimateurs « fenêtre ». Pour chaque série de mesures individuelles, la moyenne (Moy.), la somme des carrés des écarts (SCE), et les degrés de libertés (dl) sont calculés. Le tableau de données est ensuite trié sur base de la valeur de la moyenne. La variance est évaluée au sein d'une fenêtre de taille choisie sur base de la somme des carrés des écarts totale, et des degrés de liberté totaux.

Au cours d'une première étape, le jeu de données est trié, pour chaque condition expérimentale, sur base de la moyenne des valeurs d'expression pour chaque gène. A l'issue de cette première étape, une fenêtre est utilisée pour parcourir la totalité de la matrice triée. Pour chaque gène, les sommes des carrés des écarts à la moyenne sont calculés, et l'ensemble des résultats obtenus au sein de la fenêtre sont utilisés, conjointement avec les degrés de libertés associés, pour calculer la variance associée à la fenêtre définie.

L'expression mathématique qui définit le calcul de la variance associée à la fenêtre est fournie par l'équation IV.A.2.

$$\sigma_w = \frac{\sum_{i=1}^{G} SSE_i}{G(n-1)} \quad \text{(Equ. IV.A.2)}$$

où G est le nombre de gènes inclus dans la fenêtre, et n est le nombre de réplicats. SSE_i désigne la somme des carrés des écarts (*Sum of Squarred Error*) associée à chaque gène.

Nous reformulerons plus tard cette définition par une expression mathématique équivalente, pour faciliter sa comparaison avec d'autres stratégies publiées, et en dégager des informations générales.

IV.A.2.c. Caractérisation de l'estimateur fenêtre

Pour étudier l'effet bénéfique potentiel de l'effet du niveau d'expression, via l'utilisation de la fenêtre, nous pouvons utiliser les 49 mesures du jeux de données E-MEXP-231, et comparer la variance obtenue lorsqu'un nombre limité de mesures est disponible avec la variance calculée sur le jeu complet. La comparaison des deux valeurs permet de calculer l'erreur relative commise, pour chaque gène, que nous définissons par l'équation IV.A.3.

$$R.E.(i) = \frac{|sd_{W,i} - sd_{ref,i}|}{sd_{ref,i}} \quad \text{(Equ. IV.A.3)}$$

où $sd_{W,i}$ est l'estimateur fenêtre, et $sd_{ref,i}$ est la valeur attendue, estimée sur 49 réplicats. $R.E.(i)$ symbolise l'erreur relative (*Relative Error*).

Cependant, nous ignorons à ce stade quel est le nombre optimal de gènes à inclure au sein de la fenêtre, et si l'utilisation d'une grande fenêtre, qui conduirait à une variabilité exclusivement due au niveau d'expression, est préférable. En conséquence, nous devons suivre cette procédure pour plusieurs tailles de fenêtres. Nous avons choisi d'utiliser la médiane de l'erreur relative associée à l'ensemble des gènes. Cet indicateur nous permet de représenter sur un même graphe l'évolution de l'erreur commise globalement sur l'estimation de la variance. Ainsi, chaque combinaison des paramètres « nombre de gènes » et « nombre de mesures » conduit à une seule valeur représentative de l'erreur commise globalement sur l'ensemble du jeu de données évalué.

L'évolution de cet indicateur est illustrée dans la figure IV.A.3. Chaque courbe présente l'évolution de cette erreur globale (axe des ordonnées) en fonction de la taille de la fenêtre (axe des abscisses). Les différentes lignes correspondent aux résultats obtenus sur des

subsets aléatoires de 2, 3, 4, 5, 8, 10, 15 et 20 réplicats. Chaque point a été calculé sur un subset, avec une taille de fenêtre donnée. Par soucis de comparaison avec les figures précédentes, cette étude a été reproduite pour les stratégies de prétraitement utilisées précédemment.

- Le nombre de gènes présents dans la fenêtre est défini symétriquement autour du gène d'intérêt, dans un tri dépendant du niveau d'expression. Par conséquent, une fenêtre de taille n implique l'utilisation de 2n+1 gènes, et une fenêtre de taille 0 correspond à l'estimateur classique de la variance individuelle.

En assumant que la variance réelle peut être calculée sur le jeu de données complet (49 mesures), la figure IV.A.3 fournit plusieurs informations :

- L'utilisation d'estimateurs traditionnels (taille de fenêtre = 0) est associée à une erreur globale d'autant plus grande que le nombre de réplicats est petit, conformément à nos attentes.
- L'estimation de la variance sur base de plusieurs gènes permets de réduire globalement l'erreur commise lorsque la taille de la fenêtre ne dépasse pas 5 (soit 11 gènes).
- Au delà de 5 réplicats, l'estimateur traditionnel de la variance est plus approprié que l'estimateur utilisant une fenêtre. Dès lors, l'usage d'une fenêtre lorsque plus de 5 réplicats sont disponibles introduit une erreur supplémentaire, et n'est pas recommandée.

Ces observations essentielles montrent donc que la démarche envisagée repousse bel et bien les limites statistiques posées par le nombre réduit de mesures, et permet de réduire l'erreur commise globalement lors de l'estimation de la variance, en exploitant l'effet du niveau d'expression avec une fenêtre, sur base d'un petit nombre de gènes. Cependant, ces observations nous permettent empiriquement de rejeter l'hypothèse H_0, selon laquelle la variabilité est due uniquement au niveau d'expression, car la figure IV.A.3 montre qu'un nombre de gènes plus important conduit à une erreur globale plus importante, particulièrement lorsque la variabilité individuelle est estimée au départ d'un plus grand nombre de réplicats.

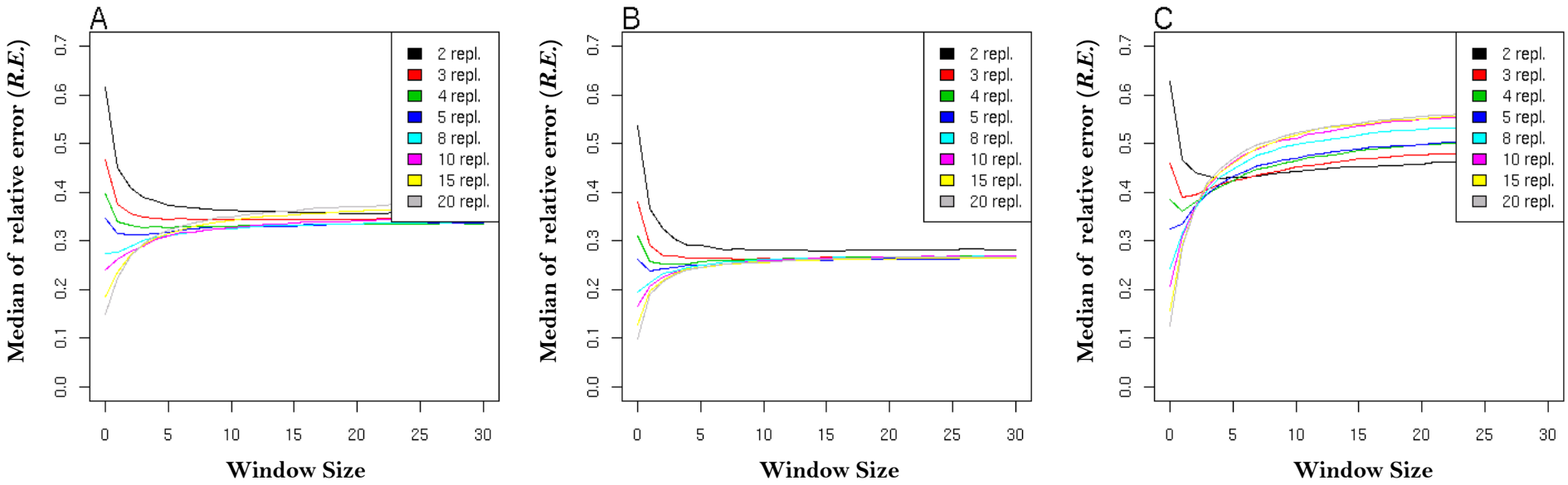


Figure IV.A.3 : Etude de l'erreur associée à l'utilisation d'une fenêtre en fonction du nombre de mesures, et de la taille de la fenêtre. Pour chaque *probeset* du jeu de données E-MEXP-231, la valeur « réelle » de l'écart-type a été calculée sur base du jeux de données complet (49 repl.). Cette valeur a ensuite été comparée à l'estimateur calculé au départ d'un jeux de données de taille réduite (de 2 à 20 réplicats), pour différentes tailles de fenêtre. L'estimateur représenté en ordonnée est la médiane de l'erreur relative commise sur l'ensemble du jeu (tous les *probesets)*. Une fenêtre de taille n implique 2n+1 *probesets*, et l'estimateur classique correspond à une fenêtre de taille 0. A: MAS 5.0 ; B: MAS 5.0 (Log 2) ; C: GCRMA. Dans tous les cas, l'utilisation d'une fenêtre de petite taille permets de réduire l'erreur commise lors de l'estimation de la variance au départ d'un jeu de données de moins de 5 réplicats. A l'inverse, l'utilisation d'une fenêtre est néfaste et augmente l'erreur commise lorsque le nombre de réplicats est supérieur à 5. Ces effets sont visibles également dans le cas de GCRMA, mais l'effet néfaste de la fenêtre sur de grands jeux de données y est plus prononcé.

La démarche optimale consiste donc à définir la taille de la fenêtre sur base du nombre de réplicats, en évitant de sur-estimer celle-ci pour éviter un effet néfaste.

Plusieurs information complémentaires se dégagent de l'observation de cette figure :

- ☞ L'utilisation d'une fenêtre de taille supérieure à 5, dans le cas d'un prétraitement avec MAS 5.0 s'accompagne d'une erreur globale constante, comparable quel que soit le nombre de réplicats utilisés. Favorablement (<5 réplicats) ou défavorablement (>5 réplicats). Il n'y a donc aucun risque à surestimer la taille de la fenêtre si le nombre de réplicats est <=5, pour ce prétraitement.

- ☞ Dans le cas de GCRMA, le comportement de l'erreur commise globalement est différent : le niveau d'erreur atteint est plus élevé si le jeu de données comporte plus de mesures. Cet effet réduit les possibilités de la méthode avec ce type de prétraitement.

- ☞ Une fenêtre de taille 2 ou 3 (5 ou 7 gènes) est optimale pour le prétraitement GCRMA, uniquement dans le cas ou le jeu comporte moins de cinq réplicats. Néanmoins, avec deux réplicats, toutes les tailles de fenêtre conviennent, et pour trois réplicats, l'utilisation de grandes fenêtres ne produit pas une erreur plus importante que pour l'estimateur classique (taille de la fenêtre = 0).

En accord avec nos attentes, des jeux de données plus petit conduisent à une erreur plus importante. L'utilisation d'une fenêtre bien choisie réduit cette erreur lorsque le jeu analysé est constitué de moins de cinq réplicats, et introduit une erreur supplémentaire dans les cas contraires.

Un point de vue différent peut également être porté sur ces observations : l'utilisation d'une fenêtre permet de réduire à deux réplicats une stratégie expérimentale prévue initialement avec cinq réplicats, en conservant une erreur similaire sur l'estimation de la variance.

Les observations illustrées ont été reproduites pour plusieurs *subsets* aléatoires, et pour d'autres jeux de données, et aboutissent aux mêmes conclusions.

IV.A.2.d. Comparaison de la variance calculée sur une fenêtre avec l'estimateur classique

Afin de comprendre le comportement de l'estimateur fenêtre, rapporté dans le paragraphe précédent, nous avons comparé celui-ci avec, d'une part, l'effet du niveau d'expression, et d'autre part, la variabilité individuelle attendue.

La figure IV.A.4 illustre la première comparaison effectuée. Nous avons utilisé une fenêtre de onze gènes (taille = 5), sur un jeu composé d'une sélection aléatoire de deux réplicats. Nous pouvons observer que la variabilité individuelle de chaque membre de la fenêtre perturbe la définition de la relation étudiée. En prenant le point de vue inverse, l'utilisation de la fenêtre centre la distribution des valeurs individuelles autour de la relation qui la relie au niveau d'expression. Lorsque la variabilité est estimée pour chaque gène sur base de la valeur caractéristique du niveau d'expression, calculée sur un petit nombre de gènes, la dispersion des points obtenues est intermédiaire entre la dispersion observée pour les valeurs individuelles, et la relation empirique présentée dans la figure IV.A.1.

Les graphiques correspondant ont été observés pour plusieurs tailles de fenêtres et pour un nombre variable de réplicats. Si le nombre de gènes utilisés pour définir la fenêtre est plus élevé, la dispersion diminue et tend vers cette même relation. L'erreur médiane commise a montré dans le paragraphe précédent qu'une fenêtre d'une taille élevée est néfaste. Les observations réalisées montrent que cet effet néfaste se produit en raison de valeurs trop proches de la relation entre le niveau d'expression et la variance. La figure IV.A.4 correspond à une situation optimale, pour laquelle la fenêtre améliore l'estimation de la variance.

La seconde comparaison réalisée, vis-à-vis de l'estimateur individuel attendu, est illustrée dans la figure IV.A.5, dans les mêmes conditions, et sur le même jeu. En comparant les graphiques obtenus pour l'estimateur classique avec ceux obtenus pour l'estimateur utilisant une fenêtre.

- ☞ L'utilisation d'une fenêtre a pour premier effet de rapprocher le nuage de points de la diagonale (qui correspond à une estimation parfaite).
- ☞ L'utilisation d'une fenêtre a pour second effet de réduire la dispersion des points (donc réduire la dispersion de l'erreur commise).

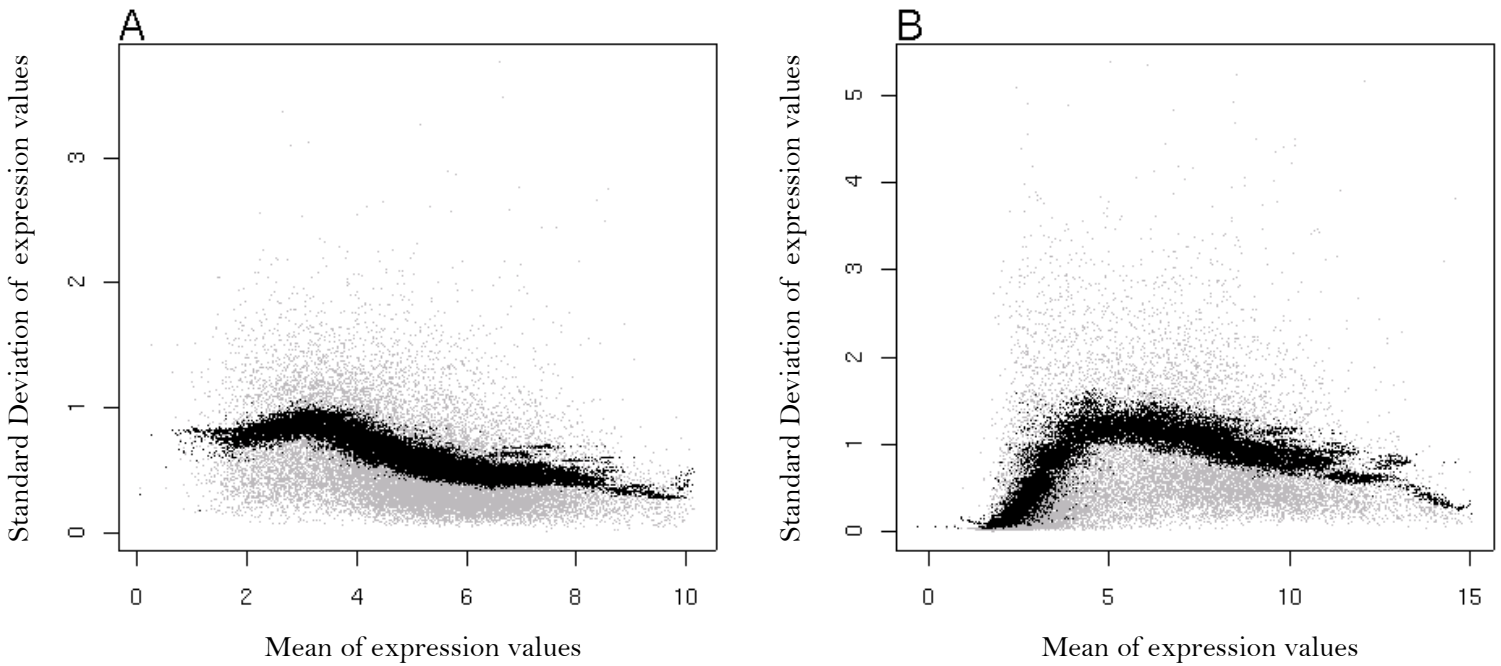


Figure IV.A.4 : Impact de l'utilisation d'une fenêtre centrée sur le niveau d'expression moyen pour estimer l'écart-type. L'écart-type, calculé sur base de la procédure classique (en gris), ou sur base d'une fenêtre (en noir) est comparé au niveau d'expression moyen pour chaque *probeset* du jeu E-MEXP-231 (A : MAS 5.0 (Log 2) ; B : GCRMA), au départ de deux réplicats. La figure montre que l'utilisation d'une fenêtre a pour effet de réduire la dispersion des points autour de sa tendance centrale associée à chaque niveau d'expression.

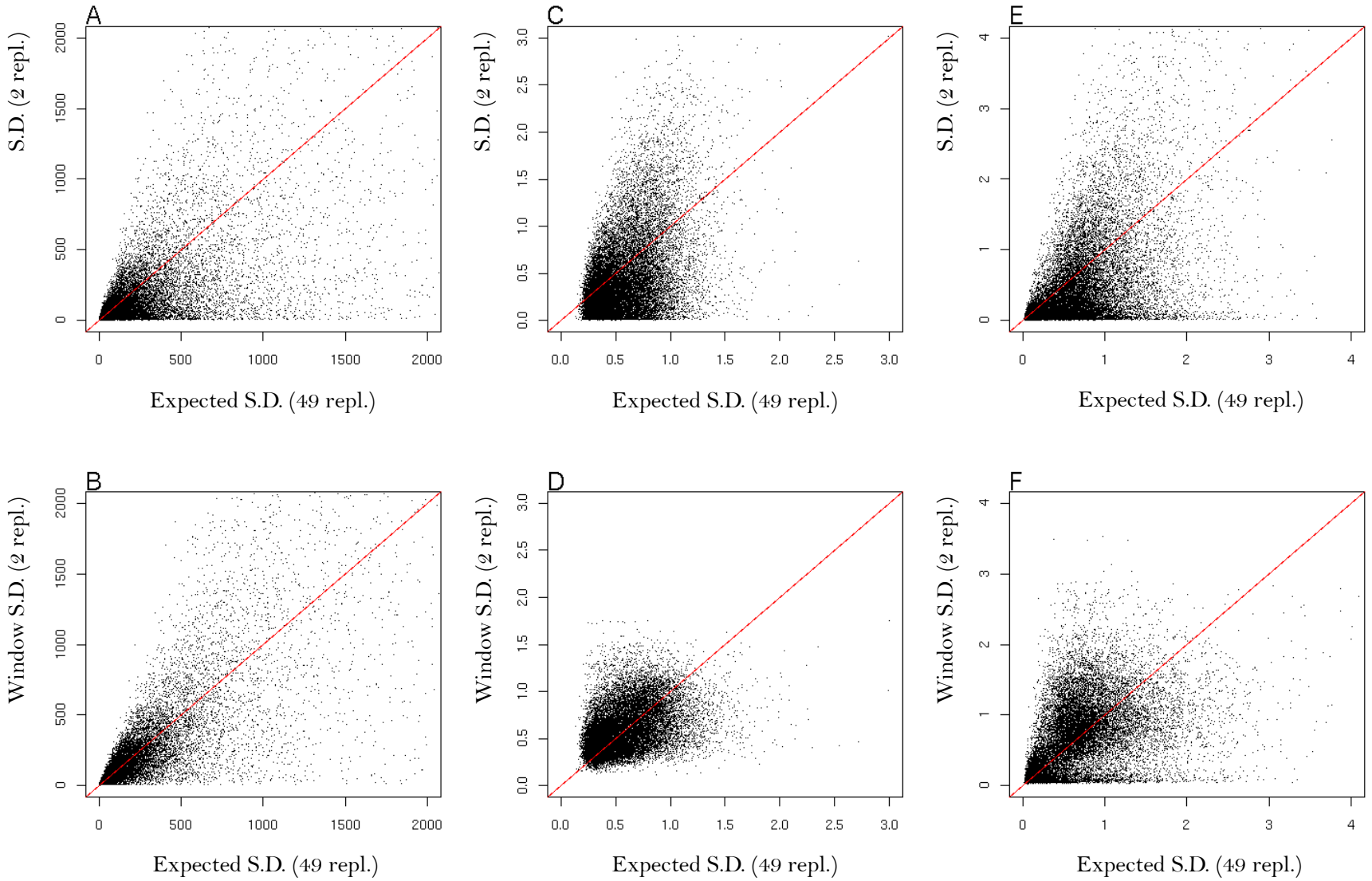


Figure IV.A.5 : Impact de l'utilisation d'une fenêtre pour estimer la variabilité individuelle. Pour chaque *probeset* du jeux de données E-MEXP-231, la valeur attendue de l'écart-type a été calculée sur base du jeux de données complet (49 mesures). Les estimateurs classiques de l'écart-type et les estimateurs « fenêtre » (taille=5) ont été calculés sur base de 2 réplicats, et comparés à la valeur attendue. A, C, E : estimateurs classiques ; B, D, F : estimateurs « fenêtre » ; A, B : MAS 5.0 ; C, D : MAS 5.0 (Log 2) ; E, F : GCRMA. Les estimateurs « fenêtre » sont plus proches de la diagonale.

- ☞ Enfin, dans le cas de GCRMA, cet effet, bien que présent, ne s'applique pas à tous les points, et un ensemble de gènes présentent un estimateur sous-estimé dans les deux cas. Cet effet pourrait être responsable du comportement différent observé pour cette méthode.

La quantification du premier effet a été présentée précédemment grâce à la médiane de l'erreur relative (Figure IV.A.3). La quantification du second effet peut être exprimée de façon similaire grâce à la MAD de l'erreur relative (*median absolute deviation*) (Figure IV.A.6).

En faisant varier le nombre de réplicats, les observations réalisées, non représentées en raison de leur nombre et de leur redondance, sont conformes aux interprétations menées sur base de la figure IV.A.3. Dans le cas de jeux de données de plus de cinq réplicats, l'estimateur classique, comparé à la valeur attendue, fourni un nuage de point centré sur la diagonale, dont la dispersion diminue à mesure que le nombre de réplicats augmente. L'utilisation d'une fenêtre avec un grand jeu de données augmente la dispersion observée, jusqu'à tendre vers une valeur constante (figuré par un nuage horizontal de points).

La figure IV.A.6 caractérise la dispersion de l'erreur commise, dans un graphique construit au départ de la déviation absolue par rapport à la médiane, pour illustrer l'évolution du second effet en fonction de la taille de la fenêtre et du nombre de réplicats. Il s'en dégage que l'usage d'une petite fenêtre avec un jeu de données de moins de cinq réplicats permets de réduire la dispersion de l'erreur commise. Au contraire, l'erreur commise est moins dispersée sur des jeux de plus de cinq réplicats, et l'usage d'une fenêtre est néfaste.

Les différentes observations rapportées dans ce paragraphe ont été reproduites tout au long de ce projet, sur plusieurs jeux de données, et fournissent les mêmes conclusions. Initialement, la procédure a été mise au point sur le jeu de données décrit par GOLUB *ET AL.* (1999) [65]. Le jeu de données E-MEXP-231 a été choisi comme exemple illustratif.

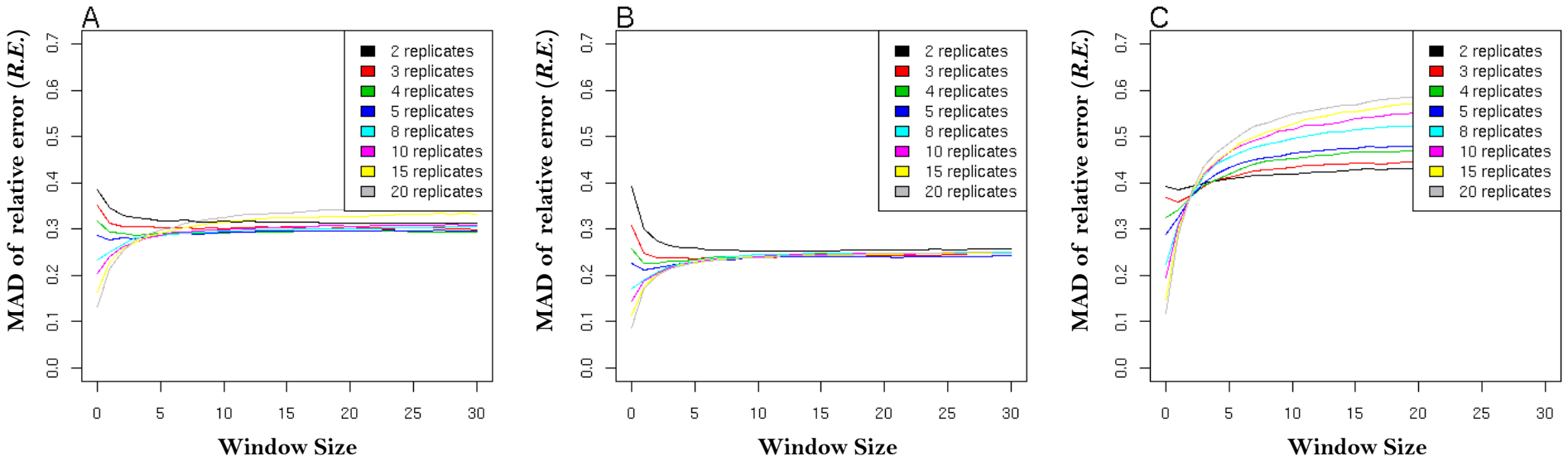


Figure IV.A.6 : Etude de la dispersion de l'erreur associée à l'utilisation d'une fenêtre en fonction du nombre de mesures, et de la taille de la fenêtre. Pour chaque *probeset* du jeu de données E-MEXP-231, la valeur « réelle » de l'écart-type a été calculée sur base du jeux de données complet (49 repl). Cette valeur a ensuite été comparée à l'estimateur calculé au départ d'un jeux de données de taille réduite (de 2 à 20 réplicats), pour différentes tailles de fenêtre. L'estimateur utilisé est la MAD (déviation absolue par rapport à la médiane) de l'erreur relative commise sur l'ensemble du jeu (tous les *probesets)*. Une fenêtre de taille n implique 2n+1 *probesets*, et l'estimateur classique correspond à une fenêtre de taille 0. A : MAS 5.0 ; B : MAS 5.0 (Log 2) ; C : GCRMA. L'utilisation d'une fenêtre de petite taille réduit la dispersion de l'erreur commise lors de l'estimation de la variance au départ d'un jeu de données de moins de cinq réplicats. A l'inverse, l'utilisation d'une fenêtre est néfaste et augmente la dispersion de l'erreur commise lorsque le nombre de réplicats est supérieur à 5. Ces effets sont visibles également dans le cas de GCRMA, mais l'effet néfaste de la fenêtre sur de grands jeux de données y est plus prononcé.

IV.A.2.e. La méthode « window t-test »

Tenant compte des observations rapportées sur l'usage d'une fenêtre, nous avons développé une nouvelle méthode dérivée du test de STUDENT, ainsi que son équivalent pour les cas d'hétéroscedasticité (correction de WELCH)[130, 143]. Notre stratégie consiste à utiliser l'estimateur « fenêtre » en lieu et place de l'estimateur classique dans le test STUDENT (BERGER ET *AL.*, 2008) [18].

La taille de la fenêtre utilisée lors de la première étape peut être spécifiée manuellement, ou, sur base de notre expérience, elle peut être automatiquement définie, via l'équation IV.A.4.

$$G = \left(\left(n_E / n_{repl} \right)_{\mathbb{N}} + 1_{(n_E /^r n_{repl}) \neq 0} \right)$$
$$W = \frac{G - 1_{(G /^r 2) \neq 0}}{2} \qquad \text{(Equ. IV.A.4)}$$

où n_E symbolise le nombre de mesures souhaitées (15 pour une bonne estimation), et $/^r$ symbolise le reste de la division entière (la fonction *modulo*). Le résultat de la division entière entre le nombre de mesures souhaitées et le nombre de réplicats définit le nombre de gènes sélectionnés, auquel on ajoute 1 si le nombre de gènes est pair (pour assurer la symétrie de la fenêtre) [18].

Cette équation a été formulée de façon à rencontrer nos attentes, et utiliser une fenêtre de taille optimale. Si l'on suppose que plusieurs gènes qui partagent le même niveau d'expression, partagent également leur variabilité, alors un grand nombre de mesures fournissent une estimation correcte de cet estimateur. Dans le cas ou le nombre de réplicats est élevé, l'estimation individuelle est plus appropriée que l'utilisation d'une fenêtre, raison pour laquelle la dépendance de la taille de la fenêtre vis-à-vis du nombre de réplicats se justifie. Nous avons choisi $n_E=15$ comme valeur afin d'éviter qu'une fenêtre trop importante soit choisie. Enfin, pour répondre à d'éventuelles limitations liées au prétraitement, nous avons combiné à cette équation la possibilité de définir une taille maximale de réplicats permettant l'utilisation d'une fenêtre. Dans pareil cas, une fenêtre de taille 0 est utilisée si le jeu de données est plus grand, et le *window t-test*, de même que plusieurs autres méthodes d'analyse dérivée du test de STUDENT, se simplifie en un test de STUDENT classique.

L'équivalent du *window t-test*, pour des variances inégales, qui implique la correction

proposée par WELCH, a également été implémentée (*window WELCH t-test*). Un troisième variant, *window mixed t-test*, effectue un test d'homogénéité des variances, et choisi en conséquence le variant approprié, pour chaque gène. Par comparaison avec les autres méthodes qui exploitent la relation entre le niveau d'expression et la variance, plusieurs éléments distinguent la méthode *window* :

- ☞ La méthode tient compte du nombre de réplicats pour définir la taille de la fenêtre, pour profiter de son effet bénéfique sur de petits jeux de données, et éviter son effet néfaste sur de grands jeux de données (>5 répl.).
- ☞ La méthode utilise une taille de fenêtre identique pour tous les gènes, contrairement à la méthode LPE, qui repose sur une fenêtre de taille variable.

IV.A.2.f. Formulations alternatives de la méthode window

La formulation générale de la méthode *window* peut être exprimée par deux autres formulations équivalentes, pour simplifier la compréhension et la comparaison avec les autres méthodes, qui sera réalisée dans le paragraphe suivant.

D'une part, ainsi que nous l'avons montré, le principe repose sur l'utilisation de l'estimateur fenêtre à la place de l'estimateur individuel. L'importance de la fenêtre est définie par sa taille, et la variance est stabilisée par intégration de données supplémentaires. Ceci peut se formuler d'une manière analogue aux autres méthodes d'estimation de la variance, sur base de deux termes, équilibrant la variance individuelle et la variance d'une fenêtre excluant le gène, pondérée par la taille de la fenêtre (Equ. IV.A.5)

$$\sigma_w^2 = \frac{\sum_{i=1}^{G} SSE_i}{G(n-1)} = \frac{SSE_g + \sum_{i=0}^{G} SSE_{i \neq g}}{G(n-1)} = \frac{(n-1)\sigma_g^2 + (G-1)(n-1)\sigma_{w-}^2}{G(n-1)} = \frac{\sigma_g^2 + (G-1)\sigma_{w-}^2}{G}$$

(Equ. IV.A.5)

où G est le nombre total de gènes de la fenêtre (dépendant du nombre de réplicats, n), et σ_{w-}^2 est la variance associée à une fenêtre centrée sur le gène d'intérêt dans un tri basé sur le niveau d'expression, mais dans laquelle le gène d'intérêt n'est pas considéré (de taille G-1). SSE_i est la somme des carrés des écarts associés à chaque gène, et σ_g^2 est

l'estimateur individuel classique de la variance.

La pondération y apparaît modulée par la taille de la fenêtre, ainsi que nous l'avons montré empiriquement.

Enfin, la méthode *window* peut être également conceptualisée différemment. L'utilisation de l'estimateur de la variance sur base d'une fenêtre en lieu et place de la variance individuelle dans le test de STUDENT est équivalent à calculer une statistique *t* classique sur des données expérimentales corrigées, de même moyenne que le jeu de départ, mais dont les données sont recalculées pour fournir une variance égale à la variance fenêtre. La correction des données peut aisément être effectuée par une transformation linéaire, formulée par l'équation IV.A.6.

$$X_W = S_W\left(\frac{X - M_X}{S_X}\right) + M_X \sim N(\mu_X, \sigma_W) \quad \text{(Equ. IV.A.6)}$$

La transformation des données d'expression correspond donc à une normalisation des données sur base de la relation empirique entre le niveau d'expression et la variabilité. Le test de STUDENT, réalisé sur ces données, fourni le même résultat que le test *window* effectué sur les données originales.

Cette dernière formulation de la correction apportée par la méthode *window* ouvre des perspectives beaucoup plus large, permettant d'étendre la méthodologie à d'autres designs expérimentaux. A titre d'exemple, la correction indépendante de données pairées permet d'appliquer la méthodologie fenêtre lorsque la procédure repose sur un test de Student adapté aux données pairées. Un test pairé permet notamment d'exploiter les données issues de biopuces de deux couleurs, ou chaque mesure d'intensité de fluorescence rouge est pairée à une mesure d'intensité de fluorescence verte. Dès lors, la transformation des données sur base de l'estimateur fenêtre peut être effectuée séparément pour les deux couleurs, et le test de *t* pairé peut être appliqué sur les données normalisées suivant le niveau d'expression. D'autre part, les tests pairés s'avèrent adaptés à certaines expérience réalisées en une seule couleur, impliquant par exemple deux personnes différentes dont des échantillons ont été prélevés pour chacune des deux conditions testées.

Enfin, cette correction des données sur base de la relation entre l'intensité et la variabilité peut-être envisagée, dans le cadre de recherches plus poussées, au niveau des données relatives aux *probes*, avant le calcul de valeurs d'expression relatives au *probesets*.

Dans les paragraphes suivants, nous présenterons les comparatifs de performances réalisés

d'une part pour distinguer les méthodes sur base de plusieurs cas de figures, et d'autre part pour valider la démarche de la méthode *window*. Avant de présenter ces évaluations, nous introduirons d'abord une comparaison théorique des méthodes envisagées, choisies pour être représentatives des méthodes disponibles. Cette démarche servira à dégager les caractéristiques communes à plusieurs méthodes d'estimation de la variance, et nous verrons que la formulation des estimateurs utilisés renforcent la validité de notre démarche. Elle facilitera d'autre part l'interprétation de l'évaluation des performances.

IV.A.3. Comparaison théorique des méthodes de correction de la variance

Les différentes méthodologies décrites dans la partie introductive de ce travail conduisent à une formulation corrigée de la statistique *t* de STUDENT. Aux premiers jours des recherches entamées dans le cadre ce cette thèse, les statistiques utilisées les plus répandues étaient le *fold change*, le test de STUDENT, le *regularized t-test (CyberT)*, et *SAM*. Le *moderated t* est apparu peu après (*Limma*). Durant ces quatre dernières années, une très grande quantité de nouvelles méthodes, toutes plus complexes les unes que les autres, sont apparues. *A posteriori*, nous savons que les meilleures performances sont atteintes par les méthodes de correction de la variance. Nous porterons notre attention particulièrement sur les méthodes sus-mentionnées, ainsi que sur la méthode *shrinkage t*, publiée en 2007, pour en dégager des enseignements communs [11, 43, 109, 124, 130, 136, 143].

La table IV.A.1 présente un résumé des développements mathématiques propres à chaque méthode, et leur formulation générale comparée. La table IV.A.2 présente quant à elle la formulation générale de l'estimateur final de la variance, et la manière dont le composant de stabilisation (*background variance* ou *null variance*) y est évalué et pondéré.

Bien que la formulation générale des différentes statistiques soit similaire, plusieurs caractéristiques essentielles les distinguent :

- ☞ la statistique sur laquelle la stabilisation s'effectue: *SAM* est la seule méthode qui corrige la déviation standard. Toutes les autres méthodes corrigent la variance [136] ;
- ☞ La pondération relative de la statistique individuelle et du terme correctif: aucune pour SAM [136], liée à la dispersion des données dans le cas de *shrinkage t* [109], des degrés de liberté dans le cas du *moderated t* [124] et du *regularized t-test* [11] ;
- ☞ Le choix arbitraire de l'estimateur utilisé pour stabiliser la variance: la médiane dans le cas de *shrinkage t* [109], la relation empirique entre le niveau d'expression et la variance dans le cas du *regularized t-test* [11] et de la méthode *window* [18], ou une procédure automatique qui utilise l'ensemble des données.

Les différentes méthodes se distinguent également par la procédure utilisée lors de l'évaluation de la significativité, et du choix des degrés de liberté associés, le cas échéant. *SAM* utilise des permutations [136], et *shrinkage t* n'évalue pas la significativité [109].

Plusieurs similitudes doivent être mentionnées entre ces méthodes, et la méthode *window* que nous avons mise au point.

- ☞ Le *regularized t-test*, à l'instar de la méthode *window*, tient compte du nombre de réplicats pour pondérer les deux composants de la variance. Cependant, celle-ci repose sur la définition d'un paramètre arbitraire, fixé à 10 [11, 18].

- ☞ La procédure du *shrinkage t*, publiée trois ans après le début de nos recherches, fait usage de la médiane, et pondère celle-ci en fonction d'une statistique intermédiaire qui quantifie la dispersion des estimateurs individuels de la variance, et de sa déviation par rapport à la valeur médiane de sa distribution. Cette démarche est donc similaire à notre étude de la dispersion des variances individuelles pour choisir la taille de fenêtre optimale sur base de l'erreur [18, 109].

- ☞ *SAM* optimise le choix de son estimateur sur la valeur associée à une dispersion minimale de la statistique d évaluée. Le même type de démarche est donc appliqué, mais est évaluée sur une statistique différente [18, 136].

- ☞ *Limma* pondère les deux composants sur base des degrés de libertés associés [124]. Ceci est vrai également pour la méthode *window*, qui utilise la taille de la fenêtre pour pondérer le terme correctif.

General formulation of the t-statistic: $t_g = \frac{D_g}{Y_g}$

Method	Equal Variances $\sigma_1^2 = \sigma_2^2$	Unequal Variances $\sigma_1^2 \neq \sigma_2^2$	Parameters $S_g = \sqrt{\frac{\Sigma_c SSE_c}{\Sigma_c df_c}}$ $SSE = \Sigma_{i=1}^{n}(x_i - M)^2$	Degrees of freedom for t distrib.
Student	$Y_g = S_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$		$S_g = \sqrt{\frac{SSE_1 + SSE_2}{n_1 + n_2 - 2}}$	$d.f.(t) = n_1 + n_2 - 2$
Welch		$Y_g = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$	$S_1 = \sqrt{\frac{SSE_1}{n_1 - 1}}$	$d.f.(t) = \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$
Regularized t-test		$Y_g = \sqrt{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}$	$S'_1 = \frac{n_{0,1} S_{0,1}^2 + (n_1 - 1) S_1^2}{n_{0,1} + n_1 - 2}$ $S_1 = \sqrt{\frac{SSE_1}{n_1 - 1}}$ $S_{\cdot,1} = \sqrt{\frac{\Sigma_{p=1}^{G} SSE_{1,p}}{N_1 - G}}$ $K = 10 = n_{\cdot,1} + n_1$ $G = 2W + 1 = 101$ (W = window size) G= #probesets $N_1 = n_1 G$ $N_1 = number\ of\ values$ in $Window$	$d.f.(t) = \frac{\frac{S_1'^2}{n_1} + \frac{S_2'^2}{n_2}}{\frac{\left(\frac{S_1'^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2'^2}{n_2}\right)^2}{n_2 - 1}}$

Table IV.A.1 - page 1 : Comparaison mathématique uniformisée de plusieurs méthodes d'analyse individuelle de l'expression différentielle et des procédures de correction de l'estimateur de la variance.

Window t-test	$Y_g = S'_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$		$S'_g = \sqrt{\frac{SSE_{0,1} + SSE_{0,2}}{N_1 - G_1 + N_2 - G_2}}$ $S_{0,1} = \sqrt{\frac{SSE_{0,1}}{N_1 - G_1}}$ $SSE_{0,1} = \Sigma_{p=1}^{G} SSE_{1,p}$ $N_1 = number\ of\ values\ \text{in}\ Window$ $G = 2\mathrm{W} + 1$ W = window size (user defined)	$d.f.(t) = n_1 + n_2 - 2$
Window Welch test		$Y_g = \sqrt{\frac{S_{0,1}^2}{n_1} + \frac{S_{0,2}^2}{n_2}}$	$S_{0,1} = \sqrt{\frac{\Sigma_{p=1}^{G_1} SSE_{1,p}}{N_1 - G_1}}$ $N_1 = number\ of\ values\ \text{in}\ Window$ $G = 2\mathrm{W} + 1$ W = window size (user defined)	$d.f.(t) = \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$
SAM d-statistic (Tusher)	$Y_g = S_0 + S_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$		$S_g = \sqrt{\frac{SSE_1 + SSE_2}{n_1 + n_2 - 2}}$	Not used (permutations).

Table IV.A.1 - page 2 : Comparaison mathématique uniformisée de plusieurs méthodes d'analyse individuelle de l'expression différentielle et des procédures de correction de l'estimateur de la variance.

Moderated-t (Limma)	$Y_g = \sqrt{\tilde{S}_g^2 v_g}$		$\tilde{S}_g = \sqrt{\frac{d_0 S_0^2 + d_g S_g^2}{d_0 + d_g}}$ $\psi'(d_0/2) = mean\left\{\frac{(e_g - \bar{e})^2 n}{n-1} - \psi'(d_g/2)\right\}$ $s_0^2 = \exp\left[\bar{e} + \psi(d_0/2) - \log(d_0/2)\right]$ $\psi, \psi' = digamma, trigamma\ functions$ $+ estimation\ de\ v_0, et\ p$	$\tilde{t}(B=0) \approx t_{d_0+d}$ $\tilde{t}(B \neq 0) \approx (1 + v_0/v)^{1/2} t_{d_0+d}$ $F(\tilde{t}_g; v_g, v_0, d_0 + d_g)$ $= p\, F\left(\sqrt{\left(\frac{v_g}{v_g + v_0}\right)} \tilde{t}_g; d_0 + d_g\right)$ $+ (1-p) F(\tilde{t}_g; d_0 + d_g)$ F(.;k) : cumulative distribution of the t distribution on k degrees of freedom. p = proportion of DE genes
Shrinkage-t	$Y_g = \sqrt{\frac{S_g^{2*}}{n}}$		$S_g^{2*} = a\, S_{median}^2 + (1-a) S_g^2$ $a = min\left(1, \frac{\Sigma_{g=1}^p \widehat{Var}(S_g^2)}{\Sigma_{g=1}^p (S_g^2 - S_{median}^2)^2}\right)$ $S_{median}^2 = median(S_g^2)$	Not used

Table IV.A.1 - fin : Comparaison mathématique uniformisée de plusieurs méthodes d'analyse individuelle de l'expression différentielle et des procédures de correction de l'estimateur de la variance.

General formulation: $V = a\ V_o + b\ V_g$

Method	Equation	Shrinkage Level	Offset Estimation
Regularized t-test	$a = \frac{n_{0,1}}{n_{0,1}+n_1-2}$ $b = \frac{n_1-1}{n_{0,1}+n_1-2}$	$V = S^2$	$S_{0,1} = \sqrt{\frac{\Sigma_{p=1}^{G} SSE_{1,p}}{N_1-G}}$ Window Estimator
SAM d-statistic (Tusher)	$a = 1$ $b = 1$	$V = S$	$S_0 = argmin\,(CV\,(MAD(d_\alpha)))$ Fudge factor minimizing CV
Moderated-t (Limma)	$a = \frac{d_0}{d_0+d_g}$ $b = \frac{d_g}{d_0+d_g}$	$V = S^2$	$s_0^2 = \exp\left[\bar{e}+\psi(d_0/2)-\log(d_0/2)\right]$ $\psi'(d_0/2) = mean\left\{\frac{(e_g-\bar{e})^2 n}{n-1}-\psi'(d_g/2)\right\}$ $\psi, \psi' = digamma, trigamma\ functions$
Shrinkage-t	$a = min\left(1, \frac{\Sigma_{g=1}^{p}\widehat{Var}(S_g^2)}{\Sigma_{g=1}^{p}(S_g^2-S_{median}^2)^2}\right)$ $b = 1-a$	$V = S^2$	$S_{median}^2 = median(S_g^2)$

Table IV.A.2 : Comparaison mathématique des procédures de stabilisation de variance employée par diverses méthodes d'analyse individuelle de l'expression différentielle. La première colonne indique le nom des méthodes. La seconde colonne indique les termes utilisés pour pondéré la variance individuelle et la variance nulle. La troisième indique la statistique sur laquelle s'applique la stabilisation de la variance. La quatrième colonne indique comment la variance nulle est calculée, et fourni l'expression des paramètres supplémentaires éventuels.

Enfin, certains choix réalisés par les auteurs, arbitraires, méritent d'être considérés avec le recul que nous offrent les observations réalisées sur la méthode *window* :

- ☞ le *regularized t-test* repose sur l'utilisation d'une fenêtre de 101 gènes, voire de la totalité de la matrice. Notre expérience montre qu'une fenêtre de cette taille se superpose avec la relation empirique observée, et représente, sur base de nos études, une valeur trop importante. Cependant, le modèle réintroduit la variabilité individuelle grâce à la pondération des deux termes [11, 18].

- ☞ La méthode *shrinkage t* utilise la médiane comme critère [109]. Par comparaison avec le *regularized t-test* et avec la méthode *window*, le terme évalué correspondrait à une procédure utilisant la totalité de la matrice comme fenêtre [11, 18]. La même remarque peut donc être formulée.

Chacune des méthodologies utilise la variance stabilisée pour calculer une statistique dérivée du t de STUDENT, dans une procédure adaptée soit pour l'égalité des variances (*SAM* [136], *moderated t* [124], *shrinkage t* [109]), soit pour l'inégalité des variances (*regularized t-test* [11], *shrinkage t* [109]).

Enfin, il est important de noter l'implication de la procédure d'évaluation de la variance stabilisée individuelle. Celle-ci peut être évaluée séparément pour chaque gène, ou, à l'autre extrême, une seule fois pour l'ensemble des gènes :

- ☞ Toutes les variances individuelles sont corrigées avec le même poids par le même terme (*SAM* [136]). Dans ce cas, l'effet individuel de stabilisation de la variance dépends de l'ordre de grandeur de la variance individuelle par rapport à l'ordre de grandeur de la constante ajoutée.

- ☞ Toutes les variances sont corrigées avec le même poids par un terme catégorisé. Dans le *regularized t-test* et dans la méthode *window*, la correction est différente pour chaque gène, en fonction du niveau d'expression. Le poids des termes est le même pour tous les gènes et dépend du nombre de réplicats [11, 18].

- ☞ Toutes les variances sont corrigées avec des poids différents pour les deux termes, et la variance *background* est la même pour tous les gènes (*moderated t* [124], *shrinkage t* [109]). La pondération des termes est calculée pour tous les gènes, et détermine la stabilisation qui est effectuée sur base d'un terme correctif unique.

L'effet final de stabilisation de la variance est donc différent pour chaque gène, agissant à

des niveaux différents (ordre de grandeur, pondération, et/ou spécificité du terme correctif), mais toutes les méthodes adoptent une approche commune qui diffère uniquement en deux points : le mode de stabilisation et la définition arbitraire des paramètres, quelle que soit la complexité du modèle mathématique envisagé.

Le développement de la méthode *window* vise à tester la capacité d'une fenêtre, sur base du niveau d'expression, à améliorer l'estimation de la variance. Les similitudes rapportées avec d'autres méthodes, en ce compris une méthode récente (*shrinkage t*) montre que toutes ces approches, d'une manière différente, reposent sur le partage d'informations entre les gènes, et fournissent les meilleures performances.

L'évaluation des performances que nous allons vous présenter au travers des prochains paragraphes nous a permis de publier la méthodologie *window*, affichant systématiquement les meilleures performances, souvent avec un niveau de performance partagé par le *regularized t-test*, *moderated t* ou *shrinkage t*, quelle que soit la stratégie d'évaluation adoptée (jeux de données simulés, jeux de données *spike-in*, jeu de données biologiques réels).

IV.A.4. Evaluation des performances

IV.A.4.a. Introduction

Parmi les méthodes d'analyse individuelle décrites précédemment, nous avons choisi de comparer les performances des méthodes dont l'essence est une correction de l'estimation de la variance. Ces méthodes, au vu des résultats publiés au sein de la communauté scientifique, sont celles qui fournissent les résultats les plus fiables, et sont les plus utilisées. A ce titre, le test de Student, utilisé comme référence, sera comparé, au cours des prochains paragraphes, au *regularized t-test* (Baldi *et al.*, *CyberT* [11]), au *moderated t* (Smyth *et al.*, Limma [124]), à la statistique *d* (Tusher *et al.*, *SAM* [136]), au *shrinkage t* (Opghen-Rhein & Strimmer [109]), ainsi qu'à la méthode *window*, développée dans le cadre de ce projet (Berger *et al.* [18]). La méthode *LPE* a également été sélectionnée, afin d'avoir un représentant supplémentaire des méthodes faisant intervenir une relation empirique entre la moyenne et la variance, et tirer globalement des conclusions sur ces méthodes (*LPE*, *regularized t-test* et *window t-test*) [11, 18, 77].

Le point le plus délicat posé par l'évaluation des performances repose sur la connaissance des résultats attendus pour l'analyse. Les auteurs des différentes méthodes proposent différents types de validation. D'une part, les évaluations quantitatives reposent sur des jeux de données simulés ou de jeux de données dénommés « *spike-in* », pour lesquels une quantité connue d'ARN est hybridée sur la puce. D'autre part, une validation qualitative est fréquemment utilisée sur base d'un jeu de données biologique réel et de l'interprétation des résultats obtenus à la lumière des connaissances externes.

Chacune de ces approches présente ses faiblesses, mais une vision globale reposant sur ces trois approches permet de tirer des conclusions globales sur les performances atteintes, et de fournir des recommandations d'usage pour les différentes méthodes.

Plusieurs types de représentations et indicateurs permettent d'évaluer les performances des résultats obtenus. Cependant, ces représentations ne sont pas équivalentes, et n'ont pas le même pouvoir de discrimination des méthodes. Le paragraphe IV.C.4 (p 229) présenté dans la troisième partie de ce travail présente une discussion relative à ce sujet.

La représentation traditionnelle des performances repose sur les courbes *ROC*. Celles-ci présentent toutefois un intérêt limité, car le nombre élevé de vrais négatifs (gènes non

impliqués) implique la superposition des courbes des différentes méthodes, et leur discrimination est quasiment impossible.

Les figures d'évaluation des performances présentées dans les paragraphes suivants reposent en conséquence sur l'utilisation des courbes *FDROC*, en accord avec les conclusions qui seront présentées dans la troisième partie. Ces figures portent en graphique la sensibilité en fonction du *FDR*. En d'autres termes, les graphiques présentés illustrent la capacité des différentes méthodes à découvrir la vérité, en fonction du prix à payer pour l'obtenir. Les performances associées aux meilleures méthodes y sont donc représentées par une courbe s'approchant le plus possible de l'extrémité supérieure gauche du graphique.

IV.A.4.b. Jeux de données simulées

Plusieurs stratégies de simulation de données ont été utilisées par les auteurs des différentes méthodes pour évaluer les performances de leurs stratégies d'analyses [118, 123]. Nous avons choisi de reproduire l'approche décrite par SHAIK & YEASIN en 2007, et de générer des données sur base de distributions normales. La stratégie originale des auteurs repose sur deux populations de gènes, respectivement identiquement et différentiellement exprimés. La première catégorie de gènes est générée sur base d'une distribution de moyenne égale à 0 et de variances individuelles générées au départ d'une distribution gamma (avec moyenne=2 et variance=2), indépendamment pour chaque gène simulé. La seconde catégorie de gènes est simulée de la même manière, mais en considérant une moyenne égale à 3 pour simuler la moitié des valeurs individuelles, et ainsi simuler une expression différentielle [123].

Afin de rendre le modèle de simulation plus complet, nous y avons apporté quelques adaptations. D'une part, parmi les « gènes » différentiellement exprimés, nous avons simulé plusieurs niveaux de réponse individuelle, avec des valeurs individuelles distribuées autour de moyennes égales à 1, 2, 3, ...10, comparées à des valeurs individuelles distribuées autour de 0. D'autre part, parmi les « gènes » identiquement exprimés, nous avons également considéré des moyennes s'étalant de 1 à 10, en complément des valeurs proches de 0 utilisées par SHAIK & YEASIN. Cette stratégie de simulation est plus proche des données biologiques réelles.

Le jeu de données simulées comporte au final 15200 gènes, parmi lesquels 10 000 ne sont

exprimés dans aucune des deux conditions, 5000 gènes sont exprimés au même niveau dans les deux conditions (répartis à différents niveaux), et 200 gènes sont différentiellement exprimés (10 gènes pour chacun des 10 niveaux de différence, en sur-expression et en sous-expression). Pour chacune des deux « conditions » simulées, trois réplicats ont été générés pour chaque « gène ».

La figure IV.A.7 illustre les performances de l'analyse de l'expression différentielle menées au départ des différentes méthodologies. Les meilleures performances sont attribuées au *regularized t-test*, suivi par la méthode *window*, démontrant ainsi la supériorité des méthodes dont la variance est modulée par une fenêtre même lorsqu'il n'existe aucune relation entre le niveau d'expression et la variabilité. Les méthodes reposant sur la modulation de la variance sur base d'un autre estimateur affichent des performances un peu plus faibles. La méthodologie *LPE*, également basée sur l'utilisation de la relation entre le niveau d'expression et la variance, fournit de moins bons résultats que les autres méthodes. Enfin, toutes les méthodologies comparées fournissent des performances largement supérieures au test de STUDENT classique.

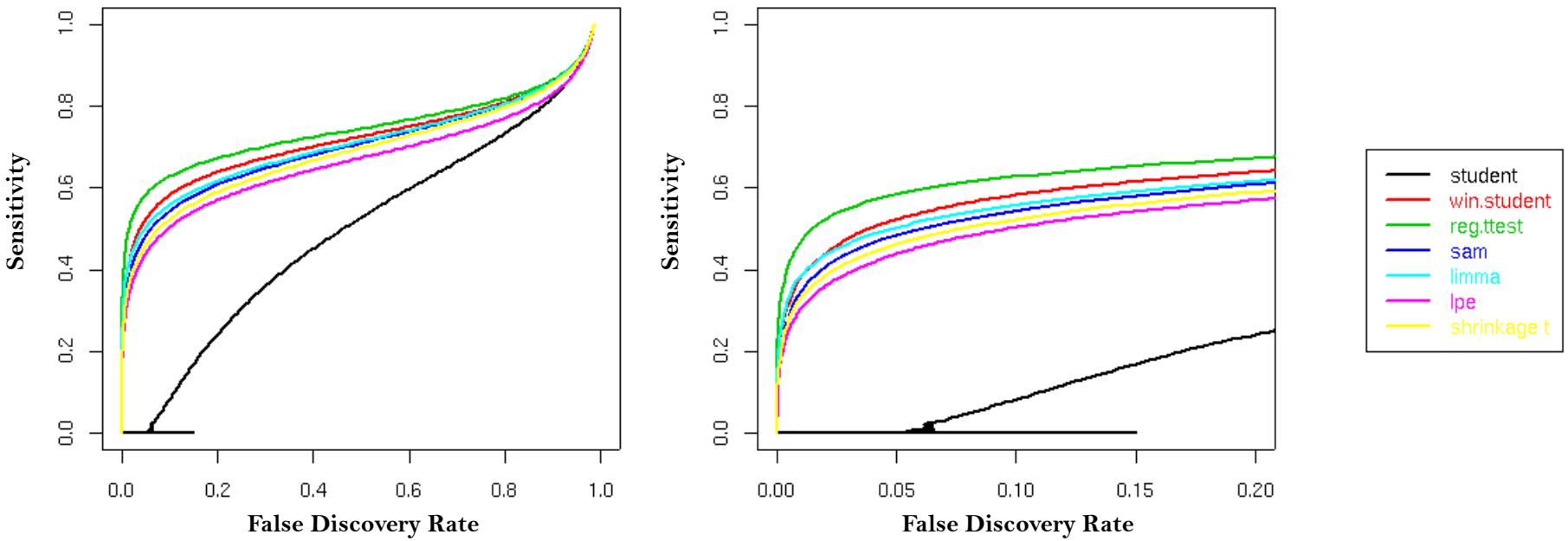


Figure IV.A.7 : Comparaison des performances de plusieurs méthodes d'analyse individuelle sur un jeu de données simulées au départ de données aléatoires. La procédure utilisée pour simuler les données est décrite dans le paragraphe VI.B.4.b. page 289. Les simulations ont été reproduites 1000 fois. Les résultats obtenus pour les 1000 simulations ont été rassemblés. Pour chaque méthode, la sensibilité (proportion de la vérité qui est détectée) est comparée au taux d'erreur *(FDR = False Discovery Rate* = taux d'erreur dans la sélection), calculée une seule fois pour l'ensemble des simulations. Le graphique illustré à droite présente une vue agrandie du graphique du gauche, pour un taux d'erreur inférieur à 20%. Le graphique illustre donc la capacité de chaque méthode à découvrir la « vérité » en fonction du « prix à payer » pour la découvrir. Les méthodes les plus performantes, plus proches du coin supérieur gauche, sont les méthodes basées sur l'utilisation de la relation entre le niveau d'expression et la variabilité, suivies par les autres méthodes méthodes basées sur le partage d'informations entre les gènes. Toutes fournissent de meilleurs résultats que le test de STUDENT.

Le gain de performance des méthodes utilisant une fenêtre s'explique intuitivement, car toutes les données ont été générées en ne considérant qu'une composante individuelle, distribuée pour tous les « gènes » simulés selon une distribution gamma dont la moyenne vaut 2. Par conséquent, l'utilisation de plusieurs gènes permet d'affiner l'estimation de la variance, et l'estimation de la variance est mieux estimée avec une fenêtre de plus grande taille. Le *regularized t-test*, utilisant beaucoup plus de mesures, estime donc mieux la variance que la méthode *window*, en raison de la distribution gamma unique utilisée pour simuler la variance de tous les gènes, contrairement aux jeux de données réels.

IV.A.4.c. Jeux de données « spike-in »

Les performances des différentes méthodes envisagées, dont la méthode *window*, développée dans le cadre de ce projet, ont été investiguées au départ de jeux de données *spike-in*. Ces jeux de données ont pour avantage d'être conçu pour évaluer les méthodologies d'analyse, en fournissant une connaissance réelle des données. Les jeux de données utilisés sont de deux types : les « carrés latins » (*Latin Square* HG-U95 [3] et HG-U133 [4]), et le *Golden Spike Experiment* (Choe *et al.*, paragraphe VI.A.2.c. page 272 [33]). Tous reposent sur l'utilisation d'ARN en quantité connue lors de l'hybridation des échantillons sur la biopuce. Dans le cas du jeu de données de Choe, cet ARN est ajouté à un échantillon lui-même composée d'un mélange d'ARN, dans le but de simuler *in vitro* un échantillon de type biologique [33].

Les figures IV.A.8 et IV.A.9 présentent les performances obtenues respectivement sur les jeux de données LS-95 et LS-133. Dans chaque cas évalué (*MAS 5.0, MAS 5.0 – log2, GCRMA*), elles présentent également une vue plus détaillée des courbes de performances pour des valeurs du FDR inférieures à 20%. L'utilisation de ce type de représentation met en évidence les méthodes les plus performantes, caractérisées par une meilleure sensibilité associée à une erreur minimale *(FDR)*.

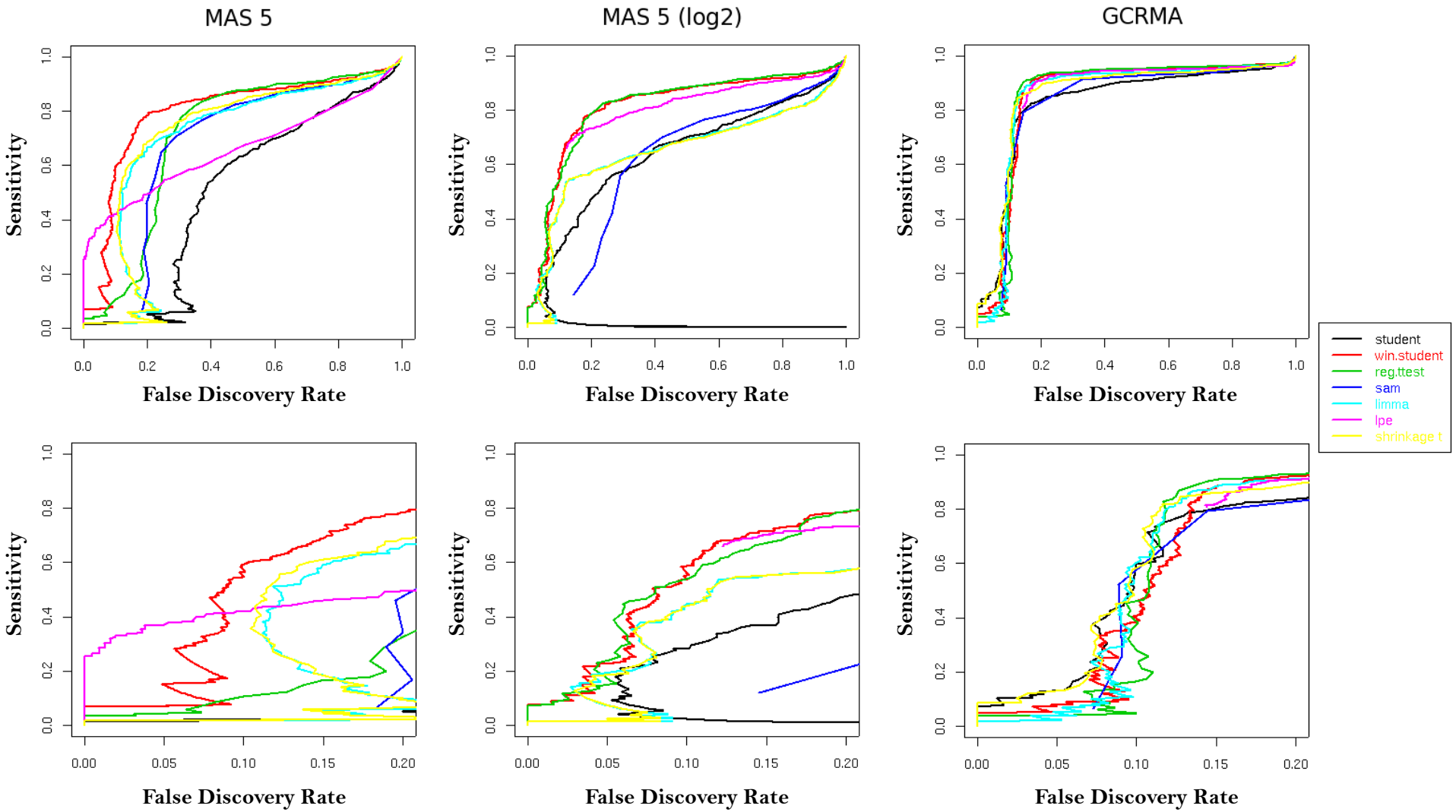


Figure IV.A.8 : Comparaison des performances de plusieurs méthodes d'analyse individuelle sur le jeu de données *Latin Square* HG-U95 (Affymetrix). Les résultats obtenus au départ de 55 analyses réalisées par comparaison deux à deux du jeu de données complet ont été rassemblés et évalués en une seule étape. Pour chaque méthode, la sensibilité (proportion de la vérité qui est détectée) est comparée au taux d'erreur (FDR = *False Discovery Rate* = taux d'erreur dans la sélection). La partie inférieure présente une vue agrandie des graphiques de la partie supérieure, pour un taux d'erreur inférieur à 20%. Les comparaisons ont été effectuées pour les prétraitements MAS 5.0 , MAS 5.0 (Log 2) et GCRMA.

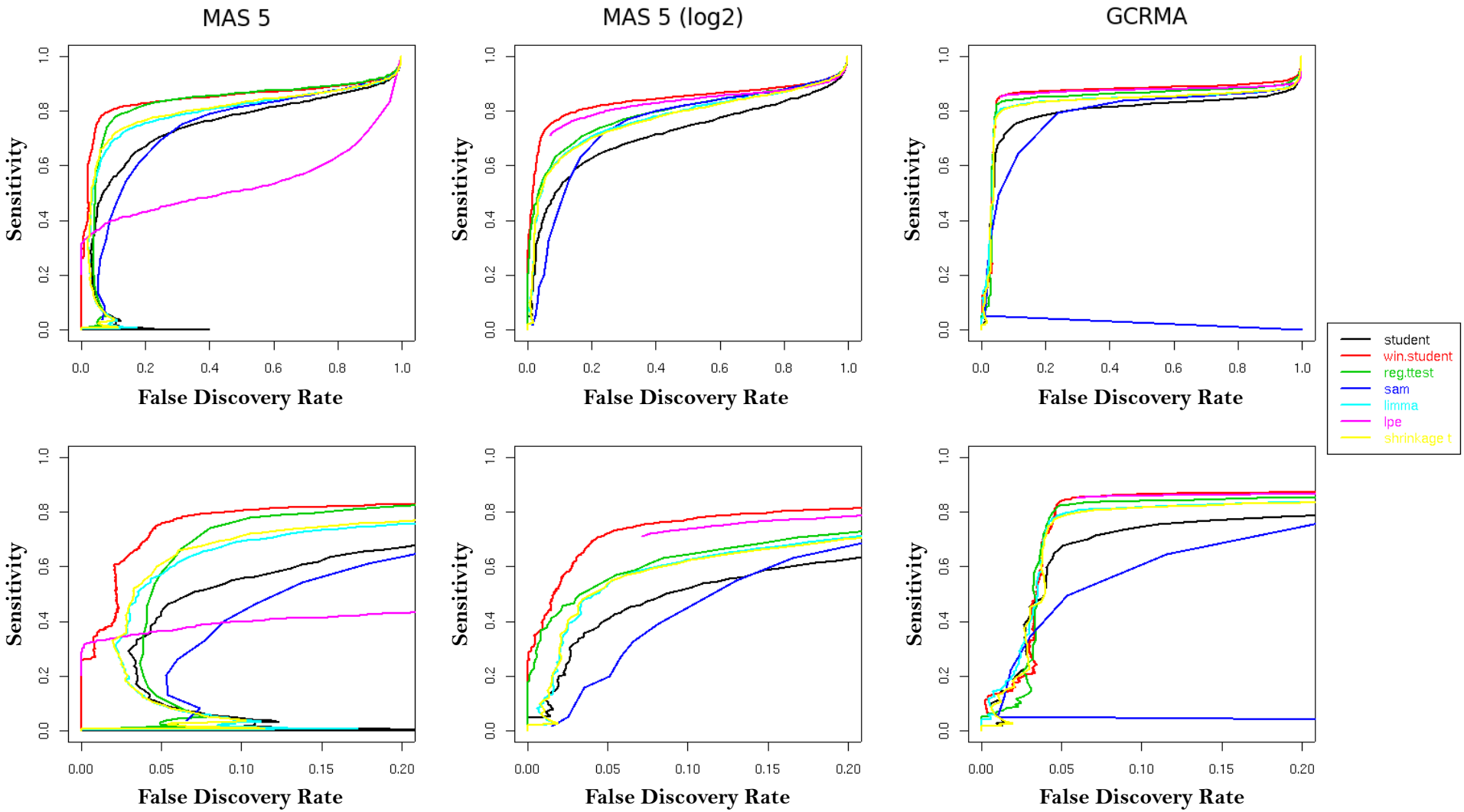


Figure IV.A.9 : Comparaison des performances de plusieurs méthodes d'analyse individuelle sur le jeu de données *Latin Square* HG-U133 (Affymetrix). Les résultats obtenus au départ de 91 analyses réalisées par comparaison deux à deux du jeu de données complet ont été rassemblés et évalués en une seule étape. Pour chaque méthode, la sensibilité (proportion de la vérité qui est détectée) est comparée au taux d'erreur (FDR = *False Discovery Rate* = taux d'erreur dans la sélection). La partie inférieure présente une vue agrandie des graphiques de la partie supérieure, pour un taux d'erreur inférieur à 20%. Les comparaisons ont été effectuées pour les prétraitements MAS 5.0 , MAS 5.0 (Log 2) et GCRMA.

L'examen des figures IV.A.8 et IV.A.9 montre que les performances sont variables selon le prétraitement utilisé, et aboutit aux observations suivantes :

- ☞ Dans tous les cas testés avec *MAS 5.0*, la méthodologie *window t-test* affiche les meilleures performances, surpassant les méthodes les plus performantes.

- ☞ Sur le jeu de donnée HG-U133A (*MAS 5.0, log2*), les performances de la méthode *window* sont de loin supérieures à toutes les autres méthodes, fournissant 80% de la vérité pour un prix à payer de 5% d'erreur au sein de la sélection. A titre comparatif, les méthodes du *regularized t-test*, *moderated t (Limma)* ou du *shrinkage t* fournissent seulement 60% de la vérité pour le même taux d'erreur, 40% avec le test de STUDENT traditionnel, et 20% avec la méthode *SAM*. En tolérant 50% d'erreur (un faux positif pour chaque vrai positif), la vérité découverte s'élève à 85% (*window*), 75-80% (*regularized t-test*, *SAM*, *Limma*, *shrinkage t*) et 70% (STUDENT *t*).

- ☞ Avec *GCRMA*, le classement des performances des méthodes sur HG-U95 n'est discriminant que pour un taux d'erreur de 15-20%, plaçant en tête les méthodes basées sur une fenêtre autour du niveau d'expression (*regularized t-test*, *window* et *LPE*) et *Limma* (*moderated t*), suivis de près par la méthode *shrinkage t*. Les méthodes les moins performantes sont le test de STUDENT et *SAM*. Pour HG-U133A, le classement attribue les meilleures performances aux méthodes basées sur une fenêtre, *LPE* et *window* surpassant le *regularized t-test*. *Limma* et la méthode du *shrinkage t* affichent des performances identiques, un peu plus faibles que les méthodes basées sur une fenêtre. Enfin, les performances les plus faibles sont attribuées au test de STUDENT et à *SAM*.

Les conclusions globales qui peuvent être portées suite à ces observations permettent de classer les méthodes en plusieurs catégories, des plus performantes aux moins performantes :

- ☞ Les méthodes dérivées du test de Student, dont la variance est modulée par un estimateur de type « fenêtre » : *window t-test* et *regularized t-test*.

- ☞ Les méthodes dérivées du test de STUDENT, dont la variance est modulée par un autre estimateur : *Limma* et *shrinkage t*.

- ☞ La méthode *SAM* et la méthode *LPE* affichent des performances plus variables,

mais globalement plus faibles que les autres méthodes. Elles affichent toutefois de meilleures performances que le test de STUDENT classique dans certaines conditions.

Ce classement est valable pour tous les cas testés, et les performances sont accrues si le prétraitement utilisé est *GCRMA*.

Outre ces principales conclusions, une série d'observations utiles peuvent être formulées :

- ☞ L'utilisation d'un prétraitement *GCRMA* conduit, pour les deux jeux de données, à de meilleures performances que le prétraitement *MAS 5.0*. En tolérant un taux d'erreur de 20% (1 faux positif pour 4 vrais positifs), 95% de la vérité est découverte au maximum, contre 80% avec *MAS 5.0*. Le départ de la courbe est stabilisé par le prétraitement *GCRMA*, favorisant les *probesets* « spikés », et les différentes méthodes se superposent, et ne peuvent être discriminées que par le niveau de sensibilité pour lequel un plateau est atteint.

- ☞ Les performances du test de STUDENT, du *regularized t-test*, de *LPE* et de *window* sont améliorées par la transformation logarithmique dans le cas du jeu HG-U95. Sur le jeu HG-U133a, seules les méthodes *window* et *LPE* affichent de meilleures performances grâce à cette étape supplémentaire. Les méthodes *Limma* et *shrinkage t* affichent la même chute de performances, et restent superposées.

- ☞ Le passage en valeurs logarithmiques semble stabiliser le départ des courbes, favorisant une évolution progressive du taux d'erreur ne faisant qu'augmenter. Le même effet est observé avec *GCRMA* (logarithmique également).

- ☞ La méthodologie *SAM* affiche un comportement particulier, avec des performances plus faibles que le test de Student pour de faibles taux d'erreur, mais atteint un plateau plus élévé si le taux d'erreur est plus important. *SAM* améliore donc les résultats du test de STUDENT, mais le prix à payer est une découverte plus tardive des *probesets* « spikés ».

IV.A.4.d. Jeux de données « Golden Spike »

L'évaluation des performances des différentes méthodes, présentées jusqu'ici au départ d'un jeu de données simulées, et des deux *latin squares* HG-U95 et HG-U133, nous a permis de classer les méthodes en trois catégories, des plus fortes aux plus faibles, et montrent en particulier la supériorité de la méthode *window*, souvent partagée avec le *regularized t-test*.

Afin de confirmer ces résultats, l'évaluation des performances a été reproduite sur le jeu de données *Golden Spike*, décrit par CHOE. Au contraire des carrés latins HG-U95 et HG-U133, celui-ci a été concu pour simuler un échantillon biologique, présentant une grande variété d'ARN hybridés, en utilisant un mélange de 3860 ARN différents, parmi lesquels 1331 ont été introduits avec des concentrations différentes pour simuler l'expression différentielle. Le taux de *probesets* « spikés » s'approche donc de 10% (1331/14010), alors que seuls 14 probesets sont spikés dans le carré latin HG-U95, et 42 dans le carré latin latin HG-U133. Néanmoins, plusieurs critiques ont été formulées par la communauté scientifique vis-à-vis de ce jeu, essentiellement dues à la très faible variabilité des *probesets* « spikés » par comparaison avec les autres. Le jeu de données est toutefois approprié pour une évaluation, à condition d'effectuer une normalisation de type *LOESS* sur les données d'expression, pour corriger cet effet [33, 115].

Les performances obtenues sur le jeu de données *Golden Spike* sont illustrées par la figure IV.A.10. Bien que les *probesets* soient tous « spikés » pour simuler une sur-expression, un test bidirectionnel a été réalisé afin de rendre plus difficile la découverte de la vérité et permettre une meilleure discrimination des méthodes. Sur ce jeu de données, le test *LPE* est le plus performant, suivi, dans l'ordre, par la méthode *window*, le *regularized t-test*, *shrinkage t*, *Limma* et enfin *SAM* et le test de STUDENT.

D'autres évaluations ont été réalisées sur le jeu de données de CHOE. Celles-ci nous ont apporté les informations suivantes, renforçant nos conclusions précédentes :

- ☞ Les performances ont également été évaluées pour un test uni-directionnel, et atteignent un niveau plus élevé, mais le classement des méthodes est conservé.
- ☞ La transformation logarithmique des données réduit considérablement les performances de la méthodes *LPE*, mais n'a aucun impact sur le classement des autres méthodes (non représenté). Les test de performances réalisés renforcent donc les conclusions exprimées sur base des carrés latins HG-U95 et HG-U133a.

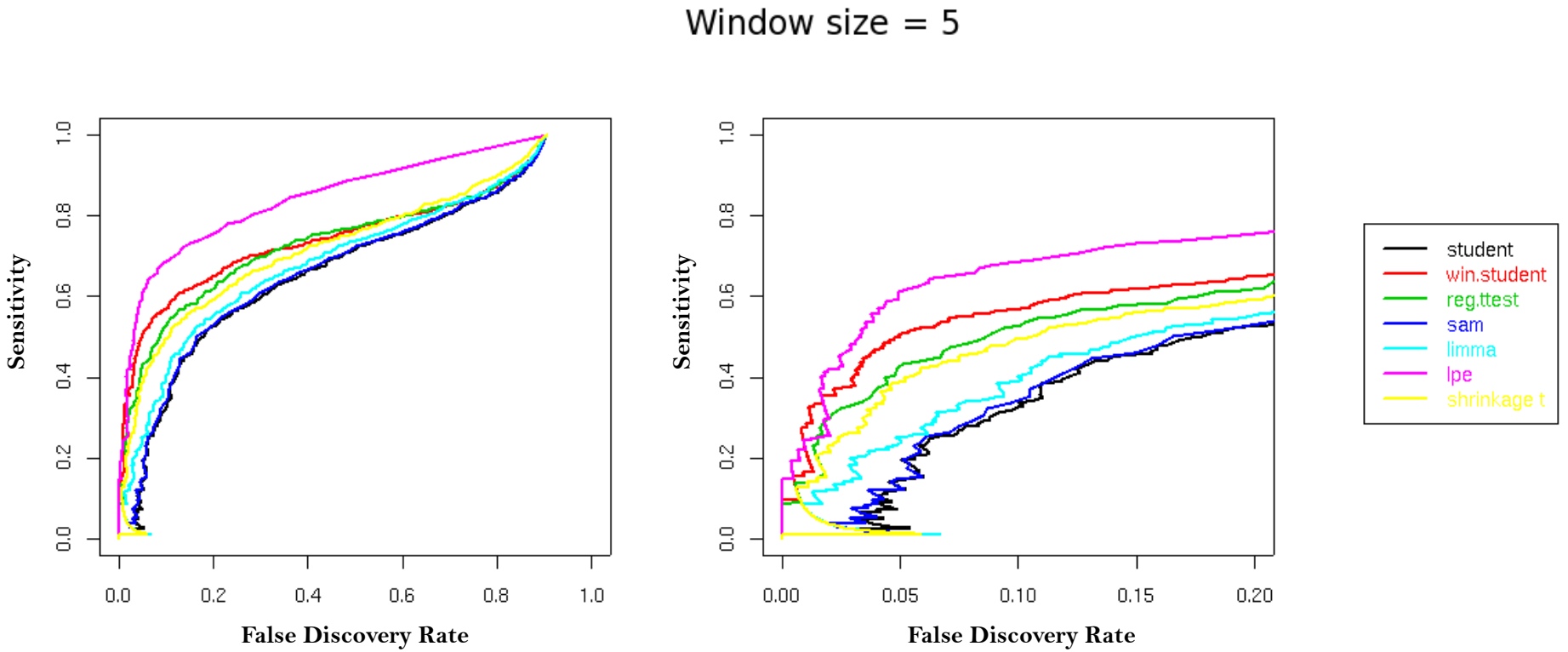


Figure IV.A.10 : Comparaison des performances de plusieurs méthodes d'analyse individuelle sur le jeu de données *Golden Spike* mis au point par Choe *et al.* Le jeux de données 10c a été utilisé, suivant les recommandations des auteurs pour une meilleure discrimination entre les méthodes. Pour chaque méthode, la sensibilité (proportion de la vérité qui est détectée) est comparée au taux d'erreur (FDR = False Discovery Rate = taux d'erreur dans la sélection). L'analyse repose sur un test bi-directionnel. Le graphique illustré à droite présente une vue agrandie du graphique du gauche, pour un taux d'erreur inférieur à 20%.

IV.A.4.e. Evaluation des performances d'une fenêtre de taille minimale

Au sein du paragraphe précédent, nous avons complété notre étude quantitative des performances des différentes méthodes, et en particulier en regard de la méthode *window*, développée dans le cadre de nos recherches. Celle-ci s'avère particulièrement efficace, atteignant les mêmes performances que le *regularized t-test*, dans presque tous les cas, les surpassant dans les autres.

La paramétrisation de la méthode *window* y a été réalisée avec une fenêtre de taille égale à 5 (11 *probesets*) pour le jeu de données de CHOE. Nous avons montré, durant la caractérisation de l'estimateur, qu'une taille de fenêtre de 3, 4 ou 5, convient parfaitement pour un jeu de cette taille. Si une fenêtre plus petite est utilisée, nous avons montré que l'effet de la fenêtre s'atténue. Qu'en est-il de la dégradation des performances de la méthode *window*, lorsqu'une taille de fenêtre minimale est utilisée ?

Pour répondre à cette question, nous avons reproduit cette évaluation des performances en utilisant une fenêtre de taille réduite (incluant respectivement 7 et 3 *probesets*). Les résultats sont illustrés dans la figure IV.A.11, par comparaison avec le *regularized t-test*, *SAM* et le test de STUDENT.

Les courbes *FDROC* présentées nous permettent de formuler les conclusions suivantes :

- ☞ le gain de performance dérivant de l'utilisation d'une fenêtre minimale (3 gènes dont le gène d'intérêt) est comparable à celui observé pour une taille optimale ;
- ☞ la méthode *window* maintient sa première place au sein du classement, suivie de près de par le *regularized t-test*.

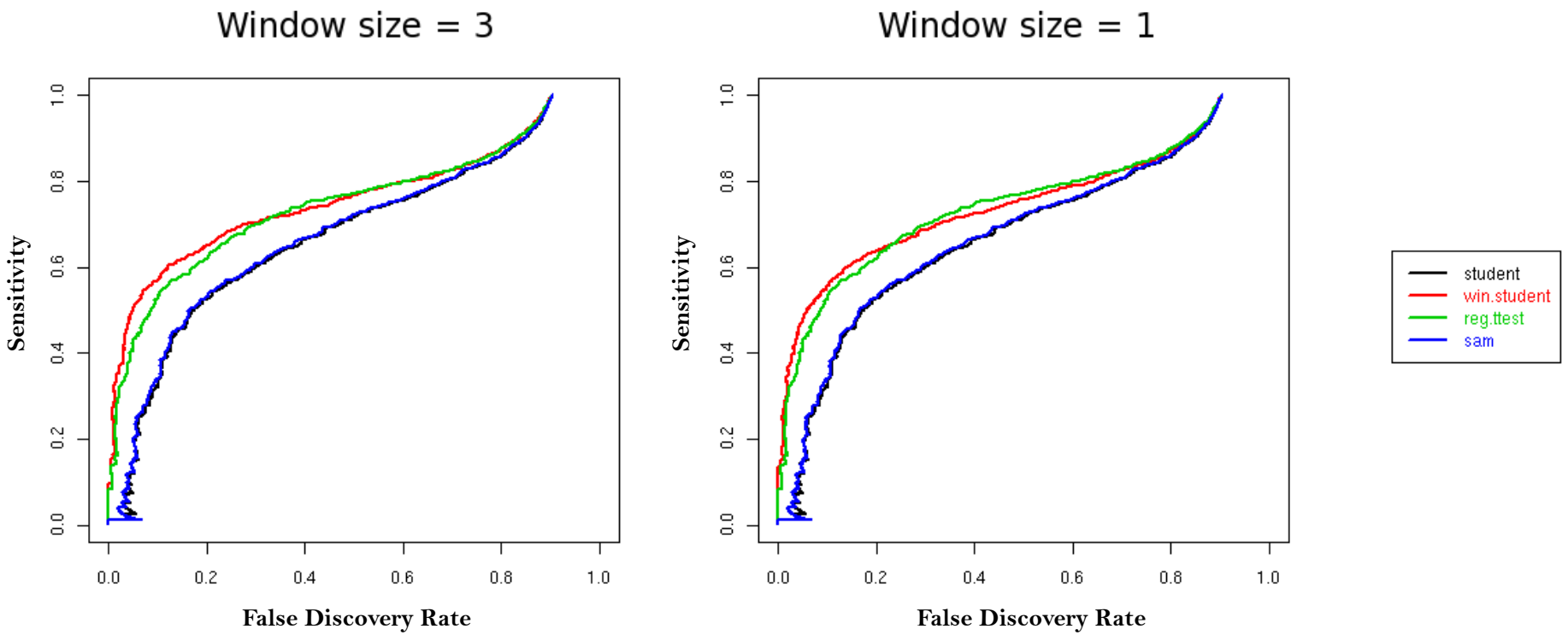


Figure IV.A.11 : Comparaison des performances de la méthode *window* sur le jeu de données *Golden Spike* mis au point par Choe *et al.* Le jeux de données 10c a été utilisé, suivant les recommandations des auteurs pour une meilleure discrimination entre les méthodes. Pour chaque méthode, la sensibilité (proportion de la vérité qui est détectée) est comparée au taux d'erreur (FDR = *False Discovery Rate* = taux d'erreur dans la sélection). L'analyse repose sur un test bi-directionnel. Les graphiques illustrés à gauche et à droite présentent les performances obtenues lorsqu'une fenêtre de taille = 3 et de taille = 1 sont utilisées (respectivement 7 et 3 *probesets*). Les méthodes *SAM* et STUDENT sont utilisées comme références.

IV.A.4.f. Jeu de données biologique

Nous avons commencé la présentation de nos résultats par l'étude de la relation empirique entre le niveau d'expression et la variabilité, avant de développer la méthode *window* sur base des études réalisées. Plusieurs méthodes d'estimation de la variance, basées sur l'échange d'informations entre les gènes, ont ensuite été comparées dans leurs fondements, et évaluées quantitativement sur base de jeux de données simulés et *spike-in*. Les performances montrent que toutes ces méthodes fournissent les meilleures performances, et les méthodes qui exploitent la relation empirique avec le niveau d'expression fournissent de meilleurs résultats, en particulier la méthode *window*, quel que soit le jeu et le prétraitement utilisé.

Pour compléter cette étude quantitative, nous avons sélectionné un jeu de données biologiques, dont l'identifiant est E-MEXP-445, comme exemple qualitatif de comparaison des résultats obtenus.

Le jeu de données E-MEXP-445, relatif aux données d'expression de monocytes humain en condition d'hypoxie et de normoxie, a été étudié. Il a été décrit précédemment par Bosco *et al.* et est disponible publiquement sur *ArrayExpress*. Les auteurs ont dressé une liste de 74 gènes connus pour leur implication dans la réponse à l'hypoxie, ainsi que d'autres gènes validés [22]. Au total, l'expression de 90 gènes (188 *probesets*) connus a été sondée pour déterminer quelles méthodes sont plus les performantes. Les méthodes qui attribuent les meilleures scores à ces 90 gènes apportent des informations supplémentaires sur les voies métaboliques et cascades de régulations impliquées, en fournissant un meilleur score pour d'autres gènes impliqués, non détectés précédemment. L'analyse du jeu de données sur base de plusieurs méthodes fourni par conséquent une vision plus détaillée des mécanismes impliqués. Sur base de nos analyses, les 100 *probesets* les plus significatifs ont été sélectionnés pour chaque méthode. Au total, 350 *probesets* ont été sélectionnés par cette procédure d'union des *top-lists*. Parmi les *probesets* sélectionnés, une recherche bibliographique a confirmé l'expression différencielle de 110 *probesets*, soit 75 gènes, dans des études en relation avec la privation d'oxygène. Pour quantifier les performances des différentes méthodes sur base d'un jeu biologique, nous avons utilisé ces 110 probesets comme estimation, incomplète, de la vérité.

Le nombre de détections validées des *probesets* au sein des 100 probesets les plus significatifs sont indiquées dans la table IV.A.3. Pour chaque méthode, la première colonne

indique le nombre de *probesets* associés aux gènes listés dans la publication originale de l'étude (Bosco *et al.* [22]), la seconde colonne indique le nombre total de *probesets* associés aux gènes validés par notre recherche bibliographique sur le sujet. Enfin, la troisième colonne indique le nombre de *probesets* pour lesquels nous n'avons trouvé aucune indication liée à la problématique de l'hypoxie. En tête de colonne, nous avons indiqué le nombre de *probesets* détectés par au moins une méthode pour illustrer la complémentarité des méthodes (union des *top-lists* individuelles).

L'examen de ce tableau aboutit à trois conclusions importantes concernant les méthodes individuelles :

- ☞ Les méthodologies qui utilisent la relation empirique entre le niveau d'expression et la variance fournissent les meilleurs résultats (*window t-test*, *regularized t-test*, *LPE test*) ;
- ☞ Les variants du test de Student et de Welch, qui utilisent la médiane et la déviation absolue à la médiane pour évaluer la statistique *t* fournissent les pires résultats en regard des connaissances actuelles des mécanismes impliqués dans la réponse à l'hypoxie. L'utilisation de la médiane et de la *MAD* comme estimateurs s'avère donc inappropriée ;
- ☞ Le nombre total de *probesets* connus et détectés par au moins une méthode est supérieur au nombre de *probesets* connus détectés par chaque méthode. Les approches envisagées conduisent donc à une complémentarité des résultats entre les méthodes individuelles.

Plusieurs gènes sont représentés plusieurs fois sur la plate-forme utilisée (HG-U133A). Cette caractéristique de multiplicité peut être utilisée pour valider les résultats, en comparaison avec la table présentée, pour énumérer le nombre de gènes validés, et obtenir une indication sur la capacité des méthodes à détecter simultanément plusieurs *probesets* relatifs au même gène. La table IV.A.4 rapporte le nombre de détections correctes des gènes, et complète donc notre analyse.

La comparaison des résultats obtenus en termes de gènes fourni d'autres observations importantes :

- ☞ Dans la catégorie des gènes connus, le nombre de gènes détectés est toujours inférieur au nombre de *probesets* détectés. Cette observation suggère l'existence d'une bonne corrélation des données propres à un même gène au sein du jeu étudié ;

- ☞ Pour les variants du test de *t* qui utilisent la médiane et la *MAD*, le nombre de gènes sélectionnés est beaucoup plus important, mais très peu sont validés ;
- ☞ Les méthodes de correction de la variance détectent davantage de gènes connus, particulièrement les méthodes qui exploitent la relation avec le niveau d'expression ;
- ☞ Le nombre de gènes détectés, mis en évidence sur base de notre recherche bibliographique, montre la validité des résultats non détectés précédemment sur ce jeu. Sur 75 gènes détectés (51 + 24), environ 1/3 a été validé récemment.

De plus, la comparaison des deux tables apporte des éléments de réponse supplémentaires :

- ☞ La quasi-totalité des *probesets* détectés par le *robust STUDENT t-test* et le *robust WELCH t-test* appartiennent à des gènes différents (97 ou 98 gènes pour 100 *probesets*) ;
- ☞ Les méthodes de correction de la variance détectent davantage de *probesets*.

La comparaison des performances en terme de *probesets* et de gènes apporte donc des éléments de réponse sur la qualités des résultats obtenus. Ces observations ne peuvent toutefois être formulées qu'en raison de la structure de la biopuce utilisée, sur base du fichier de définition standard. La redondance d'informations, qui révèle la cohérence des résultats, dérive de la multiplicité des *probesets* relatifs à un même gène. Cette multiplicité est variable pour les différents gènes, et un grand nombre de gènes ne sont représentés que par un seul *probeset*. La comparaison des deux tables serait donc différente si les gènes liés à la réponse hypoxique n'étaient représentés que par un seul *probeset*. De plus l'utilisation de fichiers de définitions alternatifs, centrés sur le gène (1 gène = 1 *probeset*) aurait rendu impossible cette observation.

Les résultats obtenus en comparant l'analyse de plusieurs méthodes sur un jeu de données biologique suggèrent donc globalement que les conclusions tirées de l'évaluation des performances réalisées sur des jeux de données *spike-in* peuvent être étendue aux cas réels. Les méthodologies *window* et *regularized t-test* s'avèrent les plus adaptées, et la cohérence des données vis-à-vis de la multiplicité des *probesets* relatifs au même gène renforce nos conclusions. Les analyses réalisées montrent également que le nombre de gènes détectés par au moins une méthode est supérieur au nombre de gènes détectés par chaque méthode. Les résultats obtenus au départ de différentes méthodes sont donc complémentaires.

A titre indicatif, la table IV.A.5 fourni la table de contingence associée à chaque méthode,

tant au niveau du *probeset* que du gène, pour une sélection des 50, 100, 150 et 200 *probesets* les plus significatifs pour chaque méthode. Ceux-ci ont été évalués préalablement à notre recherche bibliographique, sur base de la liste des gènes référencée par Bosco *et al.* [22]. Les même tendances se dégagent de l'observation de ce tableau, qui étendent donc nos conclusions pour chacun de ces seuils de sélection.

Les paragraphes suivants présentent une méthodologie mise au point pour évaluer le *consensus* des résultats. Celui-ci sera évalué sur le jeux de données *Golden Spike*, et l'analyse du jeux de données E-MEXP-445 se poursuivra ensuite, pour illustrer les capacités de celui-ci à obtenir des résultats cohérents, par comparaison avec l'intersection des listes de gènes détectés individuellement.

Probesets	Bosco et al	All known	Unknown	Total
Total	77	110	240	**350**
Student	35	44	56	**100**
Window Student	45	59	41	**100**
Welch	31	40	60	**100**
Window Welch	46	61	39	**100**
Reg. T-test	52	69	31	**100**
SAM	36	45	55	**100**
Robust Student	7	14	86	**100**
Robust Welch	8	11	89	**100**
LPE	41	57	43	**100**

Table IV.A.3 : Validation bibliographique des *probesets* sélectionnés par plusieurs méthodes d'analyse individuelle, sur base de l'analyse du jeu de données E-MEXP-445. L'énumération a été réalisée, pour chaque méthode, sur base des 100 *probesets* les plus significatifs. La première colonne énumère les *probesets* listés par les auteurs, la seconde énumère les *probesets* listés dans des études récentes, et la troisième colonne énumère le nombre de *probesets* pour lesquels aucune étude ne montre une relation avec la privation d'oxygène. Les nombres indiqués en tête de colonne caractérisent l'union des *top-lists* des différentes méthodes.

Genes	Bosco et al	All known	Unknown	Total
Total	51	75	219	**294**
Student	21	29	48	**77**
Window Student	33	43	32	**75**
Welch	18	26	55	**81**
Window Welch	33	45	33	**78**
Reg. T-test	34	45	21	**66**
SAM	22	30	47	**77**
Robust Student	5	12	85	**97**
Robust Welch	7	10	88	**98**
LPE	30	42	31	**73**

Table IV.A.4 : Validation bibliographique des gènes sélectionnés par plusieurs méthodes d'analyse individuelle, sur base de l'analyse du jeu de données E-MEXP-445. L'énumération a été réalisée, pour chaque méthode, sur base des gènes représentés par les 100 probesets plus significatifs. La première colonne énumère les gènes listés par les auteurs, la seconde énumère les gènes listés dans des études récentes, et la troisième colonne énumère le nombre de gènes pour lesquels aucune étude ne montre une relation avec la privation d'oxygène. Les nombres indiqués en tête de colonne caractérisent l'union des *top-lists* des différentes méthodes.

Informations		Positive genes:	90				Total number of genes:	13077												
		Positive probesets:	188				Total number of probesets:	22283												
Probesets	**Method**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**
	Student t-test	50	19	22064	31	169	100	35	22030	65	153	150	45	21990	105	143	200	48	21943	152	140
	Window t-test	50	27	22072	23	161	100	45	22040	55	143	150	60	22005	90	128	200	71	21966	129	117
	Welch t-test	50	20	22065	30	168	100	31	22026	69	157	150	40	21985	110	148	200	46	21941	154	142
	Window Welch t-test	50	30	22075	20	158	100	46	22041	54	142	150	63	22008	87	125	200	72	21967	128	116
	Regularized t-test	50	32	22077	18	156	100	52	22047	48	136	150	64	22009	86	124	200	73	21968	127	115
	SAM test	50	23	22068	27	165	100	36	22031	64	152	150	46	21991	104	142	200	51	21946	149	137
	Alt. SAM test	50	27	22072	23	161	100	46	22041	54	142	150	59	22004	91	129	200	69	21964	131	119
	Robust Student t-test	50	5	22050	45	183	100	7	22002	93	181	150	14	21959	136	174	200	17	21912	183	171
	Robust Welch t-test	50	6	22051	44	182	100	8	22003	92	180	150	11	21956	139	177	200	14	21909	186	174
	LPE test	50	19	22064	31	169	100	41	22036	59	147	150	57	22002	93	131	200	68	21963	132	120
	Consensus (p-value)	50	29	22074	21	159	100	49	22044	51	139	150	58	22003	92	130	200	66	21961	134	122
	Weighted Consensus (p-value)	50	26	22071	24	162	100	50	22045	50	138	150	62	22007	88	126	200	74	21969	126	114
	Consensus (rank)	50	30	22075	20	158	100	52	22047	48	136	150	61	22006	89	127	200	70	21965	130	118
	Weighted Consensus (rank)	50	31	22076	19	157	100	50	22045	50	138	150	65	22010	85	123	200	76	21971	124	112
Genes	**Method**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**	**Selected**	**TP**	**TN**	**FP**	**FN**
	Student t-test	36	11	12962	25	79	77	21	12931	56	69	124	28	12891	96	62	168	29	12848	139	61
	Window t-test	37	18	12968	19	72	75	33	12945	42	57	117	43	12913	74	47	155	46	12878	109	44
	Welch t-test	40	12	12959	28	78	81	18	12924	63	72	125	24	12886	101	66	172	29	12844	143	61
	Window Welch t-test	37	21	12971	16	69	78	33	12942	45	57	116	42	12913	74	48	154	45	12878	109	45
	Regularized t-test	34	21	12974	13	69	66	34	12955	32	56	109	42	12920	67	48	149	49	12887	100	41
	SAM test	33	12	12966	21	78	77	22	12932	55	68	122	28	12893	94	62	166	31	12852	135	59
	Alt. SAM test	36	20	12971	16	70	70	33	12950	37	57	109	42	12920	67	48	144	46	12889	98	44
	Robust Student t-test	50	5	12942	45	85	97	5	12895	92	85	144	11	12854	133	79	187	12	12812	175	78
	Robust Welch t-test	50	6	12943	44	84	98	7	12896	91	83	147	9	12849	138	81	191	9	12805	182	81
	LPE test	40	16	12963	24	74	73	30	12944	43	60	114	43	12916	71	47	149	47	12885	102	43
	Consensus (p-value)	36	20	12971	16	70	67	30	12950	37	60	108	38	12917	70	52	152	44	12879	108	46
	Weighted Consensus (p-value)	37	19	12969	18	71	67	32	12952	35	58	105	40	12922	65	50	148	49	12888	99	41
	Consensus (rank)	33	18	12972	15	72	70	33	12950	37	57	108	40	12919	68	50	148	45	12884	103	45
	Weighted Consensus (rank)	35	21	12973	14	69	67	32	12952	35	58	105	42	12924	63	48	147	50	12890	97	40

Table IV.A.5 : Evaluation du nombre de vrais et faux positifs et négatifs au sein des 50, 100, 150 et 200 *probesets* les plus significatifs du jeux de données E-MEXP-445, pour chaque méthode, sur base de la liste des gènes publiés par les auteurs de l'étude, connus pour leur implication dans la réponse hypoxique. Cette liste, bien qu'incomplète, permets de quantifier les capacités de chaque méthode à détecter les mécanismes connus.

IV.A.5. Analyse globale et *consensus*

IV.A.5.a. Introduction

Chacune des méthodes envisagées apporte une amélioration des performances de l'analyse de l'expression différentielle. Cependant les résultats sont très différents entre les différentes méthodes. Ce point de vue sera illustré et abordé sur base d'un exemple d'analyse d'un jeu de données relatif à l'hypoxie au sein des prochains paragraphes. Il ressort de cette observation que chaque méthode détecte une partie de la vérité, ce qui induit deux conséquences : les *probesets* les plus impliqués entre les conditions comparées peuvent être détectés par plusieurs méthodes, et les *probesets* un peu moins facile détecter peuvent être détectés par au moins une méthode. D'un point de vue théorique, si chaque méthode détecte une partie de la vérité, alors la combinaison des résultats individuels permet de définir des résultats plus robustes, par complémentarité ou validation croisée entre les résultats issus des différentes méthodes. Pour quantifier cette comparaison, nous avons développé une méthode d'évaluation d'un *consensus* des résultats qui vise à automatiser cette démarche, pour fournir des résultats de meilleure qualité, au départ de plusieurs méthodologies.

L'évaluation d'un *consensus* est une pratique courante en bioinformatique, dont le but est de fournir un résultat optimal sur base de la combinaison des résultats obtenus sur des tests individuels. A titre d'exemple, la méthode ESyPred3D repose sur l'utilisation de plusieurs méthodes d'alignement différentes, pour identifier une séquence similaire susceptible d'adopter une structure tridimensionnelle commune, et faciliter ainsi sa modélisation [91].

IV.A.5.b. Evaluation d'un consensus au départ de plusieurs méthodes

Dans le domaine de l'analyse de données de biopuces, le résultat fourni au cours de l'analyse de l'expression différentielle est une séquence de *p-values* associées à chacun des *probesets* représentés sur la puce. Chacune de ces *p-value*s représente la probabilité que les valeurs d'expression du *probeset* concerné soient distribuées au sein d'une même population. Les *probesets* dont la *p-value* est proche de 0 ont très peu de chance d'être associés à une même population, et ont une plus grande probabilité d'appartenir à deux populations différentes au sein des deux conditions comparées.

Les *probesets* qui appartiennent à deux populations distinctes entre les conditions comparées devraient être détectés par toutes les méthodes, avec une *p-value* proche de 0, tandis que les *probesets* non différentiellement exprimés devraient obtenir une *p-value* proche de 1 pour chaque méthode. Ces deux cas extrêmes sont accompagnés d'une grande quantité de cas intermédiaires, ou certaines méthodes détectent des *probesets* concernés, et d'autres non.

Une analyse globale de ces résultats implique donc l'estimation de la probabilité que chaque *probeset* se rapporte à une population unique pour les deux conditions, pour toutes les méthodes utilisées. Mathématiquement parlant, cette probabilité peut s'exprimer par le produit des probabilité que l'hypothèse nulle soit acceptée pour chaque méthode sélectionnée. La *p-value consensus* peut donc être calculée par le produit des *p-values* obtenues avec chaque méthode (Equation IV.A.7)

$$p_{cons}(i) = \prod_{k=1}^{n_{meth}} p_{ik} \quad \text{(Equ. IV.A.7)}$$

avec p_{ik}, la *p-value* associée au *probeset i* par la méthode *k*.

Néanmoins, sur base de cette formulation mathématique, la comparaison de listes de *p-values consensus* obtenues avec un nombre variable de méthodes conduit à une dépendance vis-à-vis du nombre de méthodes utilisées. En utilisant des opérations de passage au logarithme, le produit des probabilité peut se calculer par la somme des logarithmes des *p-values* associées à chaque méthode (Equation IV.A.8).

$$\log\left(p_{cons}(i)\right) = \sum_{k=1}^{n_{meth}} \log\left(p_{ik}\right) \quad \text{(Equ. IV.A.8)}$$

Cette procédure a été ajustée pour répondre à nos attentes (indépendance du nombre de méthodes), en calculant la moyenne des logarithmes des *p-values* méthodes-spécifiques, au lieu de leur somme (Equation IV.A.9).

$$\log\left(cons.score(i)\right) = \frac{\sum_{k=1}^{n_{meth}} \log\left(p_{ik}\right)}{n_{meth}} \quad \text{(Equ. IV.A.9)}$$

avec n_{meth}, le nombre de méthodes utilisées pour calculer le score *consensus*.

Le second problème rencontré avec cette stratégie d'évaluation du *consensus* repose sur l'échelle des valeurs de probabilités associées à chaque méthode. La figure IV.A.12 illustre la distribution des *p-values* pour chaque méthode. Chaque méthode reposant sur une stratégie d'analyse différente, les *p-values* n'y sont pas évaluées de la même manière, ce qui implique que chaque méthode fournisse une liste de *p-values* qui suivent leurs propres distributions, et l'échelle des valeurs obtenues peut être très différente d'une méthode à l'autre, même si les résultats sont proches (rangs identiques ou similaires), tel qu'observé dans la figure IV.A.12.

La procédure doit donc être adaptée en conséquence, de sorte que chaque liste de *p-values* aie le même poids au cours de l'évaluation du *consensus*. Cette question peut être abordée de différentes manières. D'une part, chaque liste de *p-value* peut être l'objet d'une correction basée sur le FDR (Bonferroni, BH, BY...). A notre connaissance, cette approche ne résout pas complètement le problème, et constitue un sujet d'étude en lui-même. Les différentes procédures existantes fournissent des résultats de différentes qualités [21, 16, 17, 50, 128, 129].

La seconde approche possible pour répondre à ce problème repose sur l'utilisation des rangs des *p-values* lors de l'évaluation du *consensus*. Un rang est assigné à chaque *p-value* au sein des listes fournies par les méthodes individuelles. En divisant ce rang par le nombre total de *p-values*, nous obtenons plusieurs listes de rangs, chacune contenant un score distribué entre 0 et 1. Cette liste de scores est ensuite utilisée en combinaison avec l'équation IV.A.9, pour évaluer le score *consensus* associé. La stratégie d'évaluation du *consensus*, adaptée pour tenir compte de la distribution variable des *p-values*, conduit à formulation de l'équation IV.A.10, où n_{ps} est le nombre total de *probesets* analysés, et r_{ik} est le rang attribué à la *p-value* du gène i par la méthode k.

$$\log\left(cons.score(i)\right) = \frac{\sum_{k=1}^{n_{meth}} \log\left(\frac{r_{ik}}{n_{ps}}\right)}{n_{meth}} \quad \text{(Equ. IV.A.10)}$$

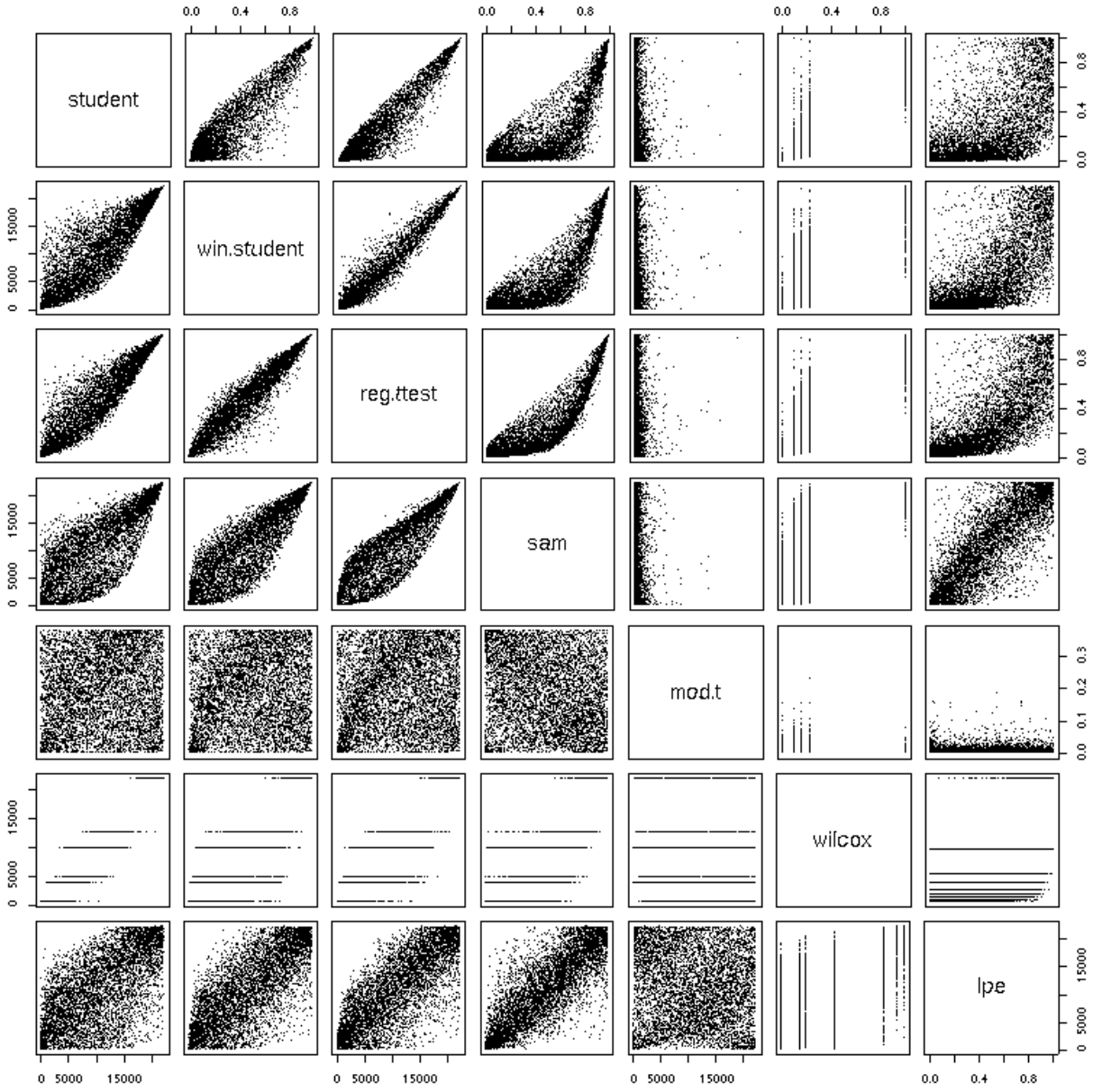


Figure IV.A.12
Comparaison de la significativité des résultats obtenus au départ de méthodes différentes. Les graphiques situés dans la partie inférieure gauche illustrent les rangs des *probesets*, et les graphiques situés dans la partie supérieure droite illustrent les *p-values* obtenues. L'analyse a été réalisée sur une comparaison de 5 échantillons choisis aléatoirement dans chacune des conditions comparées au sein du jeux de données E-MEXP-231.

Dans certaines conditions, le jeux de données étudié présente des caractéristiques qui peuvent guider le biostatisticien lors du choix des méthodes individuelles utilisées. A titre d'exemple, les méthodologies basée sur une fenêtre, ainsi que le *moderated t* et le *shrinkage t* fournissent de meilleures performances que les autres méthodes lorsque le nombre de réplicats est limité. Pour rendre la procédure d'évaluation du *consensus* plus flexible, et permettre au chercheur de tenir compte de son expérience analytique et des performances des méthodes, les équations IV.A.9 et IV.A.10 ont été adaptées pour permettre la définition de poids méthode-spécifiques. L'équation IV.A.11 fournit la formulation générale d'évaluation du *consensus*, quel que soit le score utilisé, et les poids des différentes méthodes est symbolisé par w_k.

$$\log\left(cons.score(i)\right) = \frac{\sum_{k=1}^{n_{meth}} w_k \log\left(score_{ik}\right)}{\sum_{k=1}^{n_{meth}} w_k} \quad \text{(Equ. IV.A.11)}$$

L'approche suivie pour évaluer le *consensus* des résultats vise à fournir une liste robuste de *scores*, au départ de plusieurs résultats d'analyse individuelle. Elle a été mise au point pour proposer une étape supplémentaire à l'issue de l'analyse individuelle, reposant sur la complémentarité possible des résultats issus de différentes méthodes.

La formulation du *consensus* peut toutefois s'appliquer à n'importe quelle autre thématique similaire, visant à mettre en évidence un résultat unique sur base de résultats multiples. Ainsi, le *consensus* pourrait être appliqué en méta-analyse pour combiner, gènes par gènes, les résultats obtenus sur différents jeux de données. Si ceux-ci étudient les mêmes conditions, le résultat procure une estimation des mécanismes impliqués, sur base des résultats des analyses menées séparément sur les différents jeux. A l'inverse, si les thématique étudiées diffèrent, le résultat du *consensus* se rapporte à l'intersection des analyses des jeux de données utilisés, les mécanismes communs. Enfin, pour mettre en évidence les spécificités de jeux de données différents, le *consensus* des jeux de données doit être comparé aux résultats obtenus sur chacun des jeux. Dans le contexte de l'analyse de groupe, le *consensus* présenté peut être utilisé pour obtenir une *p-value* sur base des résultats obtenus sur chaque gène membre du groupe. Une approche similaire a été utilisée précédemment par Pavlidis *et al.*, sur base des *p-values*, sans pondération, pour obtenir une valeur *consensus* par groupe de gène au départ des *p-values* individuelles [114].

L'un de nos objectifs repose sur la volonté de modélisation de la démarche analytique dans

une procédure automatisée. La comparaison des résultats obtenus sur base de différentes approches permet de fournir un résultat plus robuste. Cette qualité du *consensus* sera démontrée au sein des deux prochains paragraphes, qui reproduisent l'évaluation des performances sur base du jeux de données *Golden Spike*, et sur base du jeux de données E-MEXP-445. Les tests effectués incluent des méthodes performantes, et des méthodes peu fiables, pour en tester la robustesse. L'avantage principal de l'utilisation du *consensus* repose sur sa capacité à atteindre les mêmes performances que les meilleures méthodes, sans identification préalable ou supposition quant à ces méthodes.

IV.A.5.c. Evaluation des performances du consensus des méthodes

Au terme des comparatifs de méthodes, nous avons remarqué que les résultats fournis par différentes méthodes sont différents. Pour tirer parti de ces différences, nous souhaitons croiser leurs résultats. Dès lors, nous avons décrit une méthode simple qui nous permettra d'évaluer le *consensus* des résultats obtenus, favorisant les *probesets* détectés par plusieurs méthodes, et en défavorisant les *probesets* non détectés. Nous pensons qu'elle telle approche peut être utilisée en analyse de l'expression, pour dégager les informations majeures portées par un jeu de données.

Afin d'illustrer les performances du *consensus* des différentes méthodes, les tests de performances réalisés sur le jeu de données de Choe ont été reproduits. Au cours des différentes études menées durant ce projet, nous avons remarqué à plusieurs reprises que les performances de la méthode du produit des rangs sont peu fiables. Celle-ci a été comparée aux tests de Student, au test *window*, au *regularized t-test* et à *SAM*, sur une hypothèse unidirectionnelle (fournissant donc des résultats différents de ceux présentés précédemment). La figure IV.A.13 illustre les performances des méthodes individuelles considérées sur le jeux de données *Golden Spike*, dans un but de comparaison avec les résultats obtenus avec le *consensus* de ces méthodes.

Les quatre déclinaisons possibles du *consensus* ont été évaluées. Pour les deux approches nécessitant la définition d'un poids pour chaque méthode, ceux-ci ont été fixés comme suit : Student = 1 ; *window* = *SAM* = *LPE* = *regularized t-test* = 2 ; *rank products* = 0.5.

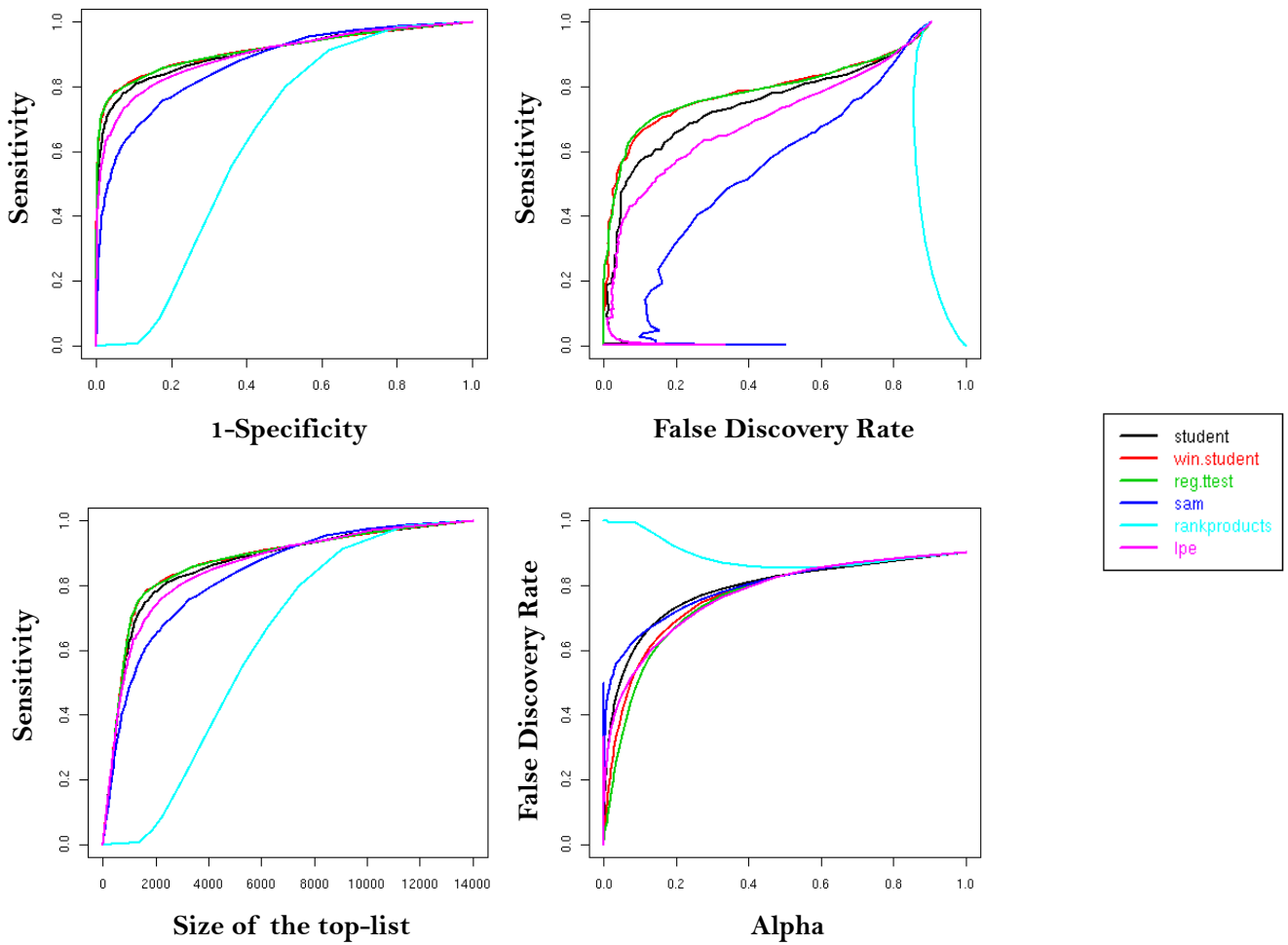


Figure IV.A.13 : Evaluation des performances de plusieurs méthodes d'analyse individuelle de l'expression différentielle sur le jeux de données *Golden Spike*.

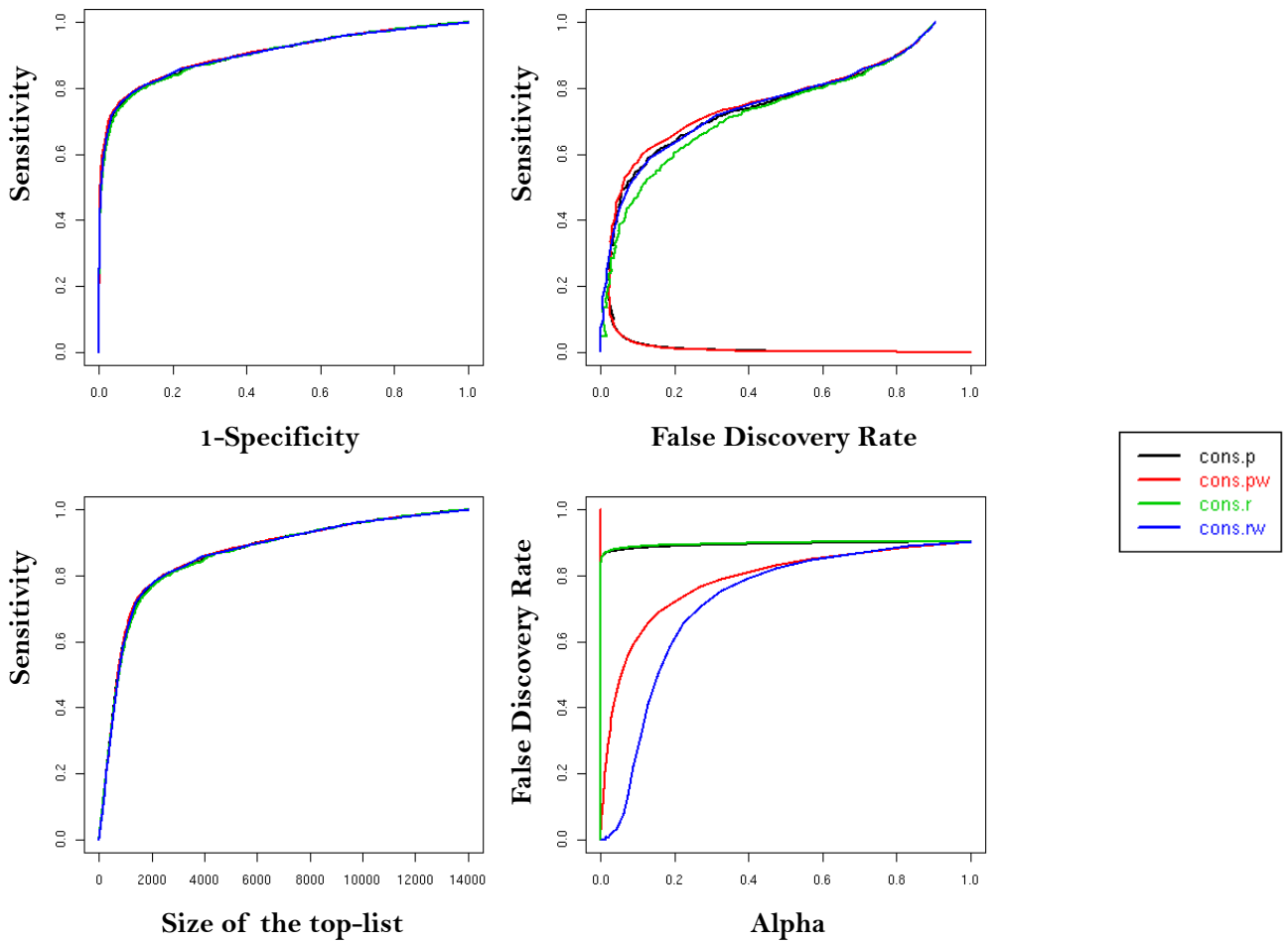


Figure IV.A.14 : Evaluation des performances des 4 variants d'évaluation du *consensus* sur base des résultats fournis par plusieurs méthodes d'analyse individuelle de l'expression différentielle, sur le jeux de données *Golden Spike*.

Les performances du *consensus* sur le jeu de données de CHOE, incluant volontairement une méthode qui fourni de mauvais résultats, sont illustrées dans la figure IV.A.14. L'examen de cette figure conduit à plusieurs observations :

- ☞ les performances atteintes par les quatre variants du *consensus* atteignent le même niveau de performance que les meilleures méthodes d'analyse individuelle inclues dans le calcul du *consensus*, malgré la présence de méthodes moins performantes (*SAM* et *rank products*) ;

- ☞ Les performances des quatre variants du *consensus* sont proches les unes des autres. En particulier, la pondération du *consensus* sur base de poids spécifiques à chaque méthode n'affecte pratiquement pas les performances du *consensus* ;

- ☞ Sur le jeu de données de CHOE, le *consensus* basé sur la *p-value* fournit de meilleurs résultats que la méthode des rangs. Ceci est du au fait que les *p-values* les plus proches de 0 ont un poids plus important, et que les meilleurs gènes se voient attribuer une *p-value* beaucoup plus petite par les meilleures méthodes. L'observation inverse a pu être obtenue sur d'autres jeux de données, lorsqu'un nombre important de faux positifs se voient attribuer une *p-value* proche de 0 (non représenté) ;

- ☞ Le score évalué par le *consensus* pondéré, sur base des rangs, fourni une meilleure approximation du *FDR* que toutes les autres méthodes (figure IV.A.14, graphique inférieur droit) ;

- ☞ La pondération manuelle du *consensus* sur les rangs, en affectant des poids plus importants aux meilleures méthodes, procure le même résultat que le *consensus* non pondéré utilisant les *p-values*.

Tous ces résultats importants montrent que le *consensus*, quelle que soit sa formulation, est capable via une procédure mathématique simple d'extraire les meilleurs résultats parmi ceux issus de méthodes adaptées et inadaptées, même lorsque nous ignorons quelle méthode est la plus adaptée (*consensus* non pondéré).

Le prochain paragraphe reproduit et complète l'exemple d'évaluation qualitative des résultats, menée, par souci de cohérence, sur le même jeux de données que celui utilisé lors de l'évaluation des méthodes individuelles.

IV.A.5.d. Evaluation du consensus des méthodes sur un jeu de données réel

Dans le cadre de l'évaluation des performances, nous avons caractérisé les résultats des méthodes d'analyse individuelle sur base d'une liste de gènes connus pour leur implication dans la thématique étudiée (l'hypoxie). L'analyse du jeu de données E-MEXP-445, à titre illustratif, a montré la supériorité des méthodes qui utilisent une fenêtre autour du niveau d'expression pour corriger l'estimation de la variance. Nous avons remarqué également que le nombre de gènes valides, détectés par au moins une méthode, est supérieur au nombre de détections correctes associées à chaque méthode. Pour tirer parti des résultats complémentaires des différentes méthodes, nous avons développé une stratégie d'évaluation du *consensus* des méthodes. Les tests de performances ont montré que le *consensus* est comparable aux méthodes les plus performantes, même en incluant une méthode qui donne des résultats de mauvaise qualité. De plus, la méthode ne nécessite aucune connaissance *a priori* des méthodes les plus adaptées au jeu de données et permet donc de dégager les meilleurs résultats au départ de plusieurs analyses individuelles.

Pour étendre les analyses du jeux de données E-MEXP-445, présentées au paragraphe IV.A.4.f. (page 144), la table IV.A.6 présente un recensement du nombre de *probesets* détectés en commun par plusieurs méthodes, et caractérise les intersections des résultats. Chaque ligne indique le nombre de méthodes utilisées pour sélectionner un *probeset*. La première colonne énumère le nombre de *probesets* représentés dans la liste des gènes publiée par les auteurs (Bosco *et al.* [22]). La seconde colonne énumère le nombre de *probesets* validés par nos recherches bibliographiques. La troisième colonne énumère le nombre de *probesets* inconnus pour la thématique de l'hypoxie.

	Bosco et al	All known	Unknown	Total
9 meth.	2	2	1	**3**
8 meth.	1	2	1	**3**
7 meth.	9	9	8	**17**
6 meth.	13	16	3	**19**
5 meth.	4	5	2	**7**
4 meth.	15	21	11	**32**
3 meth.	7	16	35	**51**
2 meth.	8	12	71	**83**
1 meth.	17	26	108	**134**

Table IV.A.6 : Comparaison du nombre de *probesets* détectés par l'intersection d'un nombre croissant de méthodes individuelles. L'analyse a été réalisée sur le jeu de données E-MEXP-445. Les 100 *probesets* les plus significatifs ont été sélectionnés pour chaque méthode. Chaque ligne du tableau correspond se réfère à tous les *probesets* représentés par l'intersection du nombre de méthodes indiqués. Les résultats sont présentés pour illustrer la quantité de *probesets* validés, soit au départ des gènes renseignés dans la publication originale (1ère colonne), soit en y incluant une recherche bibliographique récente (2ème colonne). Le nombre de gènes inconnus pour leur relation avec l'hypoxie sont repris dans la 3ème colonne. Enfin, la dernière colonne liste le nombre total de gènes détectés par intersection d'un nombre déterminé de méthodes.

	Bosco et al	All known	Unknown	Total
Total	77	110	240	**350**
Cons. P.	49	64	36	**100**
Cons. PW.	50	69	31	**100**
Cons. R.	52	64	36	**100**
Cons. RW.	50	64	36	**100**
4 cons.	45	54	23	**77**
3 cons.	2	3	8	**11**
2 cons.	6	15	7	**22**
1 cons.	3	6	9	**15**
0 cons.	21	32	193	**225**

Table IV.A.7 : Comparaison du nombre de *probesets* détectés par le *consensus* des méthodes individuelles pour les différents variants testés. L'analyse a été réalisée sur le jeu de données E-MEXP-445. Les 100 *probesets* les plus significatifs ont été sélectionnés pour chaque méthode. Chaque ligne du tableau correspond se réfère à tous les *probesets* représentés par l'intersection du nombre de méthodes indiqués. Les résultats sont présentés pour illustrer la quantité de *probesets* validés, soit au départ des gènes renseignés dans la publication originale (1ère colonne), soit en y incluant une recherche bibliographique récente (2ème colonne). Le nombre de gènes inconnus pour leur relation avec l'hypoxie sont repris dans la 3ème colonne. Enfin, la dernière colonne liste le nombre total de gènes détectés par les 4 variants du *consensus*. La partie inférieure de la table illustre les intersections des résultats fournis par les 4 variants du *consensus*, et montre une très forte corrélation des résultats entre ceux-ci.

Les observations suivantes se dégagent de l'examen de cette table :

- ☞ La catégorie de gènes la plus représentée correspond à des gènes qui ne sont pas connus pour leur implication dans l'hypoxie et ne sont détectés que par une seule méthode ;
- ☞ Il est nécessaire d'utiliser au minimum quatre méthodes pour obtenir des résultats contenant une majorité de gènes connus.

La diversité des résultats individuels est donc un aspect important de ce type d'analyse : seuls quelques gènes sont détectés simultanément par plusieurs méthodes. Ces différences entre les résultats individuels semble être essentiellement dues aux *probesets* des gènes qui ne sont pas impliqués dans la réponse à l'hypoxie.

La même approche comparative a été suivie sur base de la méthode *consensus*. La table IV.A.7 présente, en comparaison avec les tables IV.A.3, IV.A.4 et IV.A.6, le nombre de détections correctes pour chacun des variants du *consensus* (partie supérieure), ainsi que la caractérisation de leurs intersections (partie inférieure), lorsque les 100 premiers *probesets* de chaque méthode individuelle sont sélectionnés.

- ☞ La méthode *consensus*, dans ses différentes formes, fournit des résultats proches des méthodes les plus performantes, et ce malgré les mauvaises performances obtenues avec les méthodes dites « robustes ».
- ☞ Les quatre variants du *consensus* fournissent des résultats équivalents, quels que soient les poids associés aux méthodes individuelles. L'évaluation du *consensus* permet donc d'obtenir les meilleurs résultats lorsque nous ignorons quelle méthode est la plus adaptée pour analyser le jeu de données d'intérêt.
- ☞ La majorité des gènes détectés par les quatre variants sont impliqués dans la réponse à l'hypoxie (54 *probesets* connus sur 77 détectés par les 4 variants).
- ☞ De plus la plupart des gènes qui ne sont jamais détectés par l'un des *consensus* appartiennent à la catégorie des gènes « inconnus » (sur 225 *probesets* qui ne sont détectés par aucun variant du *consensus*, 193 sont inconnus dans la thématique de l'hypoxie).

Toutes ces observations confirment donc les conclusions de l'évaluation des performances de la méthode *consensus*, quelle que soit la version utilisée : les performances atteintes, en

présence de méthodes peu fiables, sont similaires aux meilleures méthodes, et ne nécessite aucune connaissance *a priori* sur les qualités des méthodes.

Sur base des informations fournies par l'union des résultats individuels, nous avons sélectionné plusieurs gènes candidats, détectés simultanément par plusieurs méthodes et/ou par plusieurs *probesets*. En raison de la cohérence des résultats et des performances obtenues par l'évaluation du *consensus*, les gènes détectés par cette approche ont également été utilisés pour définir de nouveaux candidats. La table IV.A.8 liste l'ensemble des gènes candidats, et fournit leurs scores individuels (rangs). L'expression de ces gènes détectés devrait être étudiée plus en profondeur, et caractérisée dans un environnement hypoxique. La liste complète des gènes, définie par l'union des 100 *probesets* les plus significatifs associés à chaque méthode, est fournie dans les données supplémentaires qui accompagne l'article publié sur la méthode *window* [18], et dans l'annexe I de ce document (page 309). L'annexe II page 315 reprends les références bibliographiques utilisées pour mener cette étude, et vérifier l'implication des gènes connus.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-value)	Consensus (rank)	Weighted Consensus (rank)	Probeset
DFNA5	-	67	-	66	79	-	70	-	-	79	80	74	71	203695_s_at
EGLN3	-	-	-	-	-	-	-	-	84	-	-	-	-	219232_s_at
GAS7	-	46	-	-	-	-	-	-	56	73	71	91	84	202191_s_at
	-	44	-	-	-	-	-	-	46	76	67	-	97	202192_s_at
	-	-	-	-	-	-	22	77	29	53	65	-	-	207704_s_at
	-	64	-	-	-	-	71	-	6	19	17	89	83	210872_x_at
	-	94	-	-	-	-	-	-	35	77	73	-	-	211067_s_at
HSPA6	-	-	-	-	77	-	78	-	-	-	-	-	-	213418_at
IGHG1	-	-	-	-	-	-	-	-	47	-	-	-	-	213674_x_at
	68	-	91	-	-	75	-	-	-	-	-	-	-	217039_x_at
KLHL18	57	15	-	6	-	60	-	-	-	-	81	70	55	212882_at
LASP1	21	-	24	-	-	25	-	-	-	-	-	-	-	200618_at
LGALS8	18	-	68	63	72	17	5	12	-	21	42	19	33	208934_s_at
	60	-	29	-	96	54	-	-	-	90	-	67	78	208936_x_at
	19	-	35	-	90	18	-	-	-	85	-	63	79	210732_s_at
MERTK	39	4	-	3	8	35	6	23	15	4	5	2	2	206028_s_at
	44	5	41	4	5	42	-	-	19	12	8	7	3	211913_s_at
METAP1	16	-	6	-	-	20	-	-	-	-	-	-	-	212673_at
NID1	86	-	63	-	-	81	-	-	-	-	-	-	-	202007_at
	-	-	-	-	-	-	10	18	-	-	-	-	-	202008_s_at
NR1H3	-	49	-	68	-	-	3	3	9	6	11	35	48	203920_at
OLFML2B	-	11	-	23	35	99	-	-	26	20	14	26	24	213125_at
PANX1	-	-	-	77	-	-	84	38	-	82	89	88	96	204715_at
PARVB	-	59	-	37	89	-	-	-	45	99	72	-	88	37966_at
PFTK1	-	32	-	45	-	-	-	-	-	-	94	78	70	211502_s_at
PGM1	15	9	71	17	38	14	-	-	13	10	9	12	10	201968_s_at
PPIF	92	-	42	-	-	98	-	-	-	-	-	-	-	201489_at
PRDX4	-	85	-	59	87	-	-	-	-	-	100	81	77	201923_at
RNASET2	35	56	94	54	13	32	-	-	-	58	51	39	36	217983_s_at
	36	22	-	32	14	31	-	-	59	45	38	36	29	217984_at
RXRA	55	-	28	-	-	58	-	-	-	-	-	-	-	202449_s_at
SAE2	-	-	-	-	-	-	43	52	-	92	-	90	-	201177_s_at
SDCBP	47	-	-	-	-	61	-	-	-	-	-	-	-	200958_s_at
SLCO2B1	30	2	15	1	65	29	-	-	92	27	21	13	7	203473_at
	90	-	52	-	-	89	-	-	-	-	-	-	-	211557_x_at
STK38	65	-	90	-	-	74	-	-	-	-	-	-	-	202951_at
TMEM158	-	66	-	56	-	-	-	-	10	29	26	80	74	213338_at
TMEM43	71	-	50	-	-	76	-	-	-	-	-	-	-	217795_s_at
TNS1	42	7	-	25	67	40	-	-	16	15	12	20	17	218864_at
	64	76	81	62	26	57	-	-	72	44	47	38	41	221246_x_at
	13	93	14	-	15	11	-	55	63	22	35	17	26	221747_at
	43	20	30	12	3	41	-	-	67	28	23	18	12	221748_s_at
VDAC1	84	-	75	-	-	79	-	-	-	-	-	-	-	212038_s_at
VGLL4	10	23	46	69	23	9	2	21	75	8	15	4	11	212399_s_at
	-	-	-	73	78	-	-	-	-	-	-	-	87	214004_s_at
ZNF395	-	17	-	58	74	-	-	-	27	63	45	69	58	221123_x_at

Table IV.A.8 : Liste des gènes candidats définis sur base de l'analyse du jeu de données E-MEXP-445. Les valeurs représentées font référence à la position du *probeset* étudié dans la liste des *p-values* attribuées par chaque méthode, s'il fait partie des 100 *probesets* sélectionnés.

IV.A.6. Conclusions partielles

A l'époque où nous avions initié recherches, les deux méthodes qui avaient « pignon sur rue » étaient *SAM* et le *regularized t-test*, et les publications utilisant le *fold change* et le test de Student classique étaient très répandues [11, 43, 130, 136, 143].

Nous avons formulé l'hypothèse que l'analyse individuelle peut être améliorée en partageant de l'information entre les gènes sur base d'un critère valide de regroupement des gènes. En utilisant le seul critère connu alors pour sa validité empirique, la relation entre le niveau d'expression et la variance, nous avons étudié l'utilisation de plusieurs gènes, définissant la « fenêtre », et en avons tiré des enseignements concernant son utilisation optimale en relation avec le nombre de mesures disponibles et le nombre de gènes considérés.

Nous avons ensuite présenté la méthode *window t-test*, et l'avons constamment comparé avec les procédures disponibles et publiées récemment, pour en dégager des démarches communes, qui valident notre approche : la correction de la variance, qui peut se formuler sous une forme universelle pour toutes les méthodes, et qui repose sur le partage d'informations entre les gènes, sur base de diverses procédures de pondération de l'information partagée. En particulier, nous avons montré que la méthode *shrinkage t*, publiée en 2007, rejoint notre démarche de pondération basée sur l'étude de la dispersion de l'erreur individuelle [109].

Les évaluations de performances réalisées sur 3 types de jeux de données (simulés, *spike-in* et biologique) ont ensuite démontré que la méthode *window* procure des résultats de qualité similaire ou supérieure aux méthodes les plus performantes, dans tous les cas testés, lorsque le nombre de réplicats disponible est limité (<=5). En particulier, le même niveau de performance est atteint lorsqu'une fenêtre de taille minimale est utilisée.

Nous avons ensuite proposé l'ajout d'une nouvelle étape, l'évaluation du *consensus* (testé sur base d'une formulation mathématique simple), suivant une approche d'analyse globale des résultats issus de plusieurs méthodes pour en combiner les avantages. Les tests d'évaluation réalisés ont montré que la méthode *consensus*, évaluée au départ de méthodes sub-optimales, classiques et optimisées, fournit un résultat de bonne qualité, sans nécessiter l'identification de la meilleure méthode dans un contexte donné. Des résultats complémentaires ont montré également que le *consensus* stabilise l'évolution des indicateurs de performances lors du parcours des résultats.

Les résultats obtenus avec la méthode *window t-test*, le *regularized t-test* et la méthode *consensus*, lors de l'analyse d'un jeu de données réel a montré l'aptitude de ces méthodes à détecter un plus grand nombre de gènes connus pour leur implication dans la problématique étudiée.

Tous ces résultats nous ont permis de publier la méthode *window t-test* dans la revue *Central European Journal of Biology* [18].

En conclusion, l'utilisation de plusieurs gènes pour estimer la variance est une démarche validée par nos travaux et par d'autres études, et le niveau atteint dépend du critère utilisé pour regrouper les gènes, et pour assurer la pondération des informations individuelles et partagées. Les objectifs visés par notre démarche sont donc atteints et nous encouragent à mener plus avant ces études, en utilisant d'autres critères, définis sur base de connaissances biologiques, pour définir la « fenêtre » utilisée.

La seconde partie du chapitre Résultats présentera les premières recherches que nous avons menées en ce sens, dans le but d'améliorer les performances de l'analyse de groupes de gènes connus. Nous pensons qu'il sera possible, grâce aux outils développés dans cette seconde partie, d'étudier plus avant les connaissances actuelles pour en identifier des critères utilisables en combinaison avec la méthode *window t-test*.

La troisième partie du chapitre Résultats présentera la modélisation de la stratégie d'analyse optimale, grâce aux enseignements tirés de notre démarche, et son automatisation au sein du *package* logiciel PEGASE.

IV.B. Analyse de l'expression différentielle de groupes de gènes

Résumé

Ce chapitre présente les recherches menées sur la thématique de l'analyse de l'expression de groupes de gènes. Par « analyse de groupes », nous désignons les études qui s'intéressent à la manière dont un groupe fonctionnel connu évolue suivant les conditions de l'expérience. Ceci les distingue des études de coexpression qui recherchent des corrélations dans l'expression de gènes, et des méthodes de *clustering* qui visent à découvrir de nouveaux groupes (*group discovery*) sur base des conditions de l'expérience.

En initiant ce projet, les méthodes d'analyse principalement utilisées reposaient sur les méthodes de sur-représentation, qui ne tiennent compte que des gènes dont la modification d'expression est la plus significative. Cette analyse est cependant tronquée d'un point de vue biologique, car quelques gènes dont l'expression est peu modifiée peuvent néanmoins induire une effet métabolique considérable.

Le postulat à l'origine du projet repose sur l'utilisation de la procédure *ANOVA-2* pour améliorer les performances de l'étude de groupes de gènes. D'un point de vue méthodologique, ce postulat correspond à l'étude multivariée de la variabilité des gènes d'un même groupe pour déterminer l'effet dû à la condition expérimentale.

Nous proposons une méthode bidirectionnelle dérivée, *FAERI*, qui repose sur deux étapes simples de transformation des données. La statistique *F* de *FAERI* ne suit pas la distribution associée à la statistique de FISHER classique. Nous proposons d'évaluer la significativité au départ de permutations, ou de données aléatoires.

Les comparatifs de performances ont été réalisés en regard de la taille des groupes, de la direction de la réponse, de la proportion de membres impliqués, et de la corrélation entre les gènes. Les évaluations réalisées sur des données aléatoires, uni- ou multivariée(s) montrent que *FAERI* présente des performances supérieures aux autres méthodes, et trouve davantage de groupes recherchés. L'*ANOVA-2* classique est toutefois plus appropriée lorsque les membres sont tous soit sur-, soit sous-exprimés.

L'analyse de trois jeux de données relatifs à l'hypoxie confirme que les méthodes découvrent des groupes reliés aux mêmes voies métaboliques. Les méthodes bi-directionnelles découvrent davantage de groupes associés plusieurs voies de signalisation et à certaines pathologies où la réponse hypoxique a été observée (cancers). Les résultats de *FAERI* sont les plus corrélés entre les trois jeux de données, quelle que soit la source de définition des groupes. Le résultat des analyses incite donc à utiliser *FAERI* et l'*ANOVA-2* pour optimaliser l'analyse de l'expression par groupe de gènes.

IV.B.1. Introduction

Les données d'expression mesurées sur des *microarrays*, peuvent être étudiées à différents niveaux (analyse individuelle, analyse de groupes de gènes, études de coexpression, clustering) [67]. L'analyse de groupes de gènes reliés par un critère biologique connu ou observé présente un intérêt tant fondamental que clinique ou appliqué. A titre d'exemple, les membres d'une même voie métabolique peuvent être étudiés simultanément pour déterminer son implication dans les conditions étudiées, pour étudier le mécanisme d'action d'un médicament, ou pour diagnostiquer une pathologie... [147]. Ce dernier cas de figure est fréquemment utilisé sur base de groupes appelés « signatures », déterminés pour diverses pathologies grâce aux études de coexpression [142] et de *clustering* [9, 89]. L'étude de groupes de gènes implique deux démarches complémentaires :

- ☞ la recherche de groupes : les données empiriques sont utilisées pour mettre en évidence de nouveaux critères, pour découvrir de nouveaux groupes (*group discovery*). Cette démarche implique différentes approches, telles que le *clustering* des données (définition de signatures)[54, 65, 72, 133], l'étude des séquences régulatrices [23, 28, 29, 73, 88, 111, 113, 138] ...

- ☞ l'analyse de groupes : les définitions de groupes de gènes connus sont utilisées pour guider l'analyse des données empiriques, pour déterminer si l'expression des groupes est influencée par les conditions des expériences réalisées [21, 6, 7, 13, 32, 44, 45, 48, 53, 63, 64, 69, 74, 80, 82-85, 87, 89, 90, 93, 101, 104, 106, 107, 114, 122, 131, 132, 135, 140, 150, 151].

Les recherches que nous présentons dans cette seconde partie s'inscrivent dans cette seconde démarche. Notre motivation initiale vise à répondre à la question Q_0 : « parmi les groupes connus, quels sont ceux au sein desquels les membres sont exprimés différemment entre les deux conditions comparées ? ». Cette question implique plusieurs hypothèses relatives à la nature des groupes définis: les gènes membres peuvent être exprimés à des niveaux différents, ils peuvent être sur- ou sous-exprimés, faiblement ou fortement. Nous nous intéresserons également à la question Q_U : « quels sont les groupes qui sont globalement sur-exprimés ou sous-exprimés » (groupes uni-directionnels). Cette question est incluse dans la question Q_0, mais ne s'intéresse pas aux groupes qui contiennent simultanément des membres sous-exprimés et sur-exprimés, identifiés par la question Q_B (groupes bi-directionnels).

Nous dresserons tout d'abord une liste comparative des principales méthodes actuellement disponibles, en regard de leurs procédures, de leurs spécificités, et de leur capacité à répondre aux questions Q_0, Q_U et Q_B. Nous montrerons, par cette comparaison, que plusieurs hypothèses différentes sont envisagées par les différentes méthodes, qui apportent donc des résultats complémentaires, et que seules certaines méthodes en deux étapes sont utilisables pour répondre à la question Q_0. Ces méthodes n'exploitent toutefois pas la totalité des informations disponibles, car elles résument les données associées à chaque gène par une seule valeur, considérée comme représentative.

Nous proposons de répondre à la question Q_U sur base d'une procédure *ANOVA-2* (deux critères de classification croisés) et à la question Q_0 en adaptant la procédure afin de permettre l'évaluation de l'implication des groupes de gènes connus, quelle que soit la direction de la réponse individuelle des membres, indépendamment du niveau d'expression.

La statistique F^* évaluée par la méthode proposée, dénommée *FAERI*, n'est pas distribuée selon la distribution classique de FISHER en raison des adaptations proposées. Nous présenterons deux approches différentes d'évaluation de la distribution nulle de F^*, basées sur des données aléatoires ou sur des permutations d'échantillons, en accord avec l'hypothèse du test.

Les deux méthodologies *ANOVA-2* et *FAERI* seront ensuite comparées aux autres méthodes disponibles en terme de performances. Deux stratégies de simulations complémentaires permettront de comprendre quelles sont les propriétés des groupes les mieux détectés par chaque méthode, sur base des critères les plus déterminants : direction de la réponse, corrélation entre les membres, proportion de membres impliqués, et nombre de membres. Nous montrerons, par ces simulations, que la procédure *ANOVA-2* est la plus appropriée pour l'étude de groupes uni-directionnels, que la méthode *FAERI* fournit les meilleurs résultats lors de l'étude de groupes bi-directionnels, et qu'elle s'avère appropriée pour l'étude simultanée de tous les types de groupes testés.

Nous fournirons ensuite un exemple d'analyse relatif à la thématique de l'hypoxie (privation d'oxygène), et montrerons d'un point de vue qualitatif que les résultats obtenus sont cohérents, que les groupes mis en évidence par différentes méthodes sont complémentaires et conformes aux connaissances actuelles des mécanismes impliqués. La comparaison des listes de gènes montrera l'intérêt d'étudier simultanément les deux types de groupes (avec deux méthodes adaptées). Enfin, la corrélation des résultats entre plusieurs expériences sera illustrée pour chaque méthode.

IV.B.2. Comparaison théorique des méthodes existantes

Plusieurs méthodes d'analyse de groupes, correspondant à diverses stratégies d'analyse, ont été présentées dans la partie introductive de ce travail. Parmi elles, les méthodes de sur-représentation visent à identifier les groupes de gènes connus dont les membres présentent la plus forte différence entre les conditions, définis sur base d'un seuil arbitraire placé en aval de l'analyse individuelle. Elles ne considèrent donc que la *top-list* obtenue. Cette approche, bien qu'intéressante, correspond davantage à une annotation qui facilite l'interprétation de l'analyse individuelle, en lui donnant un sens biologique. Plusieurs gènes appartenant à une même voie métabolique fournissent des arguments pour la perturbation de la voie métabolique concernée. Les méthodes de sur-représentation présentent cependant un intérêt limité et ne seront pas considérée dans notre étude, car des groupes de gènes dont tous les membres sont affectés faiblement sont ignorés, malgré l'impact biologique qu'ils peuvent représenter [6, 7, 13, 32, 45, 82-85, 101, 151].

Pour faciliter la compréhension du lecteur, nous proposons de reformuler les hypothèses testées en tenant compte des propriétés des groupes biologiques analysés. Chaque groupe est caractérisé par la réponse de ses membres. Les deux critères qui nous semblent les plus importants, en accord avec les études rapportées, sont la direction et la corrélation entre les membres. De plus, le niveau d'expression est variable selon les gènes, qui ne sont pas identiquement distribués (et n'ont pas le même poids au cours de l'analyse).

Q_0 : *« Quels sont les groupes dont l'expression diffère entre les expériences réalisées ? »*

Q_{cor} : *« Quels sont les groupes dont les membres présentent une réponse corrélée ? »*

Q_{uni} : *« Quels sont les groupes sur- ou sous-exprimés ? »*

Q_{bidir}: *« Quels sont les groupes dont les membres sont différentiellement exprimés ? »*

Q_{int}: *« Quels sont les groupes au sein desquels la réponse des membres est variable ? »*

Dans le cadre de nos recherches, nous voulons répondre à la question générale Q_0, qui englobe tous les cas de figures possibles. La table IV.B.1 fournit un aperçu non exhaustif, mais représentatif, des différentes méthodes disponibles. Elles sont comparées sur base des hypothèses testées (3ème colonne), des données utilisées (4ème colonne), de la statistique employée pour caractériser le groupe (5ème colonne) et du mode d'évaluation de la significativité (6ème colonne). Les autres colonnes sont fournies à titre indicatif.

Auteurs	Année	Hypothèse	Données utilisées	Statistique de groupe	Significativité	Nom	Propriétés	Statistique individuelle
Mootha et al	2003	Compétitive	statistique individuelle	*ES (Running Sum)*	Permutations d'échantillons	GSEA	Hybride	rapport signal/bruit (SNR)
Subramanian et al	2005	Compétitive	statistique individuelle	*ES (Running Sum)*	Permutations d'échantillons	GSEA	Hybride,asymetrique	corrélation individuelle (r)
Keller et al	2007	Compétitive	statistique individuelle	*ES (Running Sum)*	Modèle théorique compétitif	variant GSEA	Hybride,symétrique	*
Effron & Tibshirani	2007	Compétitive	statistique individuelle	*ES (Running Sum)*	Permutations d'échantillons	GSA	*Restandardization*	*
Pavlidis et al	2004	Compétitive	statistique individuelle	log[p(g)]=mean{log[p(i)]}	Permutations de gènes		Dépend de la taille du groupe	coefficient de corrélation de Pearson
Tian et al	2005	Compétitive	statistique individuelle	Moyenne (pondérée)	Permutations de gènes		standardisation	t de Student
Tian et al	2005	Auto-suffisante	statistique individuelle	Moyenne (pondérée)	Permutations d'échantillons		standardisation	t de Student
Kim & Volsky	2005	Autosuffisante	statistique individuelle	moyenne	Distribution normale	PAGE	Théorème Central Limite	Fold-change
Effron & Tibshirani	2007	Auto-suffisante	statistique individuelle	moyenne	Permutations d'échantillons	GSA	Uni-directionnel + *Restandardization*	t de Student
Effron & Tibshirani	2007	Auto-suffisante	statistique individuelle	*maxmean*	Permutations d'échantillons	GSA	Uni-directionnel (sous groupe directionnel) + *Restandardization*	t de Student
Effron & Tibshirani	2007	Auto-suffisante	statistique individuelle	*absmean*	Permutations d'échantillons	GSA	Uni-directionnel (valeur absolue) + *Restandardization*	t de Student
Dinu et al	2007	Autosuffisante	Données d'expression	somme(d^2)	Permutations d'échantillons	SAM-GS	détermination de s0	statistique d de SAM
Goeman et al	2004	Auto-suffisante	Données d'expression	Q(g)=moyenne(Q(i))	Permutation/Gamma /Asymptotic	GlobalTest	P(Y|X)	Q(i)
Mansmann & Meister	2005	Auto-suffisante	Données d'expression	F	Permutations d'échantillons	GlobalAncova	P(X|Y)	
Berger et al	*non publié*	*Auto-suffisante*	*Données d'expression*	*F*	*F de Fisher*	*ANOVA-2*	*Uni-directionnel*	
Berger et al	*non publié*	*Auto-suffisante*	*Données d'expression*	*F**	*Permutations d'échantillons / données aléatoires*	*FAERI*	*Bi-directionnel*	

Table IV.B.1
Comparaison des propriétés de plusieurs méthodes d'analyse de l'expression différentielle de groupes de gènes. Les méthodes sont groupées en plusieurs catégories. Les parties supérieures et inférieures du tableau listent respectivement les méthodes compétitives et auto-suffisantes. Les méthodes sur-lignées en gris reposent sur une procédure en deux étapes. Les méthodes inscrites sur fond blanc reposent sur une analyse dite « globale », qui utilise les données d'expression en une seule étape pour définir la statistique de groupe, sur base de modèles multivariés. Enfin, l'ANOVA-2 et FAERI, présentées dans ce travail, sont indiquées en italique.

La première distinction entre les méthodes repose sur la manière dont la significativité du groupe est évaluée, en regard de l'hypothèse testée. Les méthodes qui étudient l'hypothèse compétitive visent à classer les groupes en répondant à la question Q_{comp} : « Pour quels groupe de gènes la réponse observée est-elle différente des autres définitions de groupes possibles ? » [53, 80, 106, 114, 131, 132]. Cette démarche, pour évaluer correctement la significativité, est associée à des permutations de gènes (définition aléatoire de groupes de même taille) [64].

A l'inverse, d'autres méthodes, dites « auto-suffisantes » (*self-contained*) évaluent la significativité sur base de permutations d'échantillons au sein d'un même groupe. L'attribution de la *p-value* s'effectue en répondant à la question Q_{Self} : « La distribution particulière des mesures observées au sein du groupe est-elle différente entre les conditions définies en regard d'une définition aléatoire des conditions comparées ? ». Chaque groupe sera donc évalué par rapport à l'ensemble des réponses possibles propres au groupe. La distribution nulle peut être évaluée pour chaque groupe grâce à des permutations d'échantillons (définition aléatoire des conditions comparées) [44, 53, 63, 74, 87, 104, 132].

Enfin, les méthodes *GSEA* et ses variants présentent une stratégie hybride, dont les résultats sont plus délicats à interpréter, car ce sont des méthodes compétitives, mais elles évaluent la significativité sur base de permutations d'échantillons [80, 106, 131].

La question Q_{Comp} vise à identifier les groupes qui présentent une réponse plus forte que les autres groupes, et non à identifier les groupes qui sont simplement « différents » entre les deux conditions (Q_0). Les méthodes auto-suffisantes sont plus appropriées pour apporter une réponse à la question Q_0, en évaluant cette différence uniquement sur base de la définition des conditions, indépendamment des autres groupes.

Le second critère de distinction des méthodes repose sur la définition de la statistique utilisée pour quantifier la réponse observée. Le premier groupe de méthodes repose sur une analyse en deux étapes. Les données individuelles des gènes membres sont tout d'abord caractérisées par une statistique individuelle. Celle-ci est ensuite utilisée à la place des données, pour chaque membre, afin d'évaluer une statistique associée au groupe [53, 80, 87, 106, 114, 131, 132]. Le second groupe de méthodes, dites « globales », effectuent l'analyse en une seule étape sur base de l'ensemble des données d'expression du groupe (*GlobalTest*, *GlobalAncova*) [44, 63, 104]. *SAMGS* est un cas particulier des méthodes globales dont la formulation finale est similaire aux méthodes en deux étapes [44].

La question Q_0 implique plusieurs questions liées à la nature des groupes testés. Ceux-ci peuvent être caractérisés par plusieurs critères, dont trois seront pris en considération :

- ☞ le niveau d'expression : l'étude des différences d'expression individuelles implique des gènes répartis à des niveau d'expression différents. Il est important d'en tenir compte pour éviter que des gènes fortement exprimés masquent la différence observée pour des gènes faiblement exprimés. La plupart des méthodes en deux étapes peuvent être utilisées avec une statistique individuelle indépendante du niveau d'expression, ainsi que la méthode *SAMGS*. A titre d'exemple la statistique *t* de Student et *d* de *SAMGS* caractérisent la différence d'expression, quel que soit le niveau d'expression individuel [44].

- ☞ la direction de la réponse individuelle : les différences d'expression individuelles peuvent révéler une sous- ou sur-expression des gènes entre les conditions testées. L'indépendance du test vis à vis de la direction implique la prise en compte de toutes ces différences, en évitant que des groupes mixtes ne présentent un effet global nul (50% sous-exprimés et 50% sous-exprimés). En réponse à ce critère, les méthodes en deux étapes peuvent être adaptées pour utiliser la valeur absolue d'une statistique individuelle uni-directionnelle (*t* de Student [130]), son carré (*SAMGS* [44]), une statistique de groupe qui combine les deux sous groupes directionnels (*absmean*), ou représentative du sous-groupe directionnel qui fourni la plus forte réponse (*maxmean*) [53]. Au sein des méthodes globales, *SAMGS* et *GlobalTest* sont indépendante de la direction, car l'hypothèse nulle est testée respectivement sur base d'une statistique quadratique (d^2) et de la variabilité de la réponse (τ^2) [44, 63].

- ☞ la corrélation entre les membres : les gènes sont susceptibles de présenter une réponse corrélée, ou de répondre de façon variable. La question Q_0 couvre ces deux possibilités. Pour y répondre correctement, il convient donc de caractériser la différence totale observée pour le groupe indépendamment de la répartition des contributions individuelles. Les méthodes en deux étapes basées sur une somme ou une moyenne, ainsi que *SAMGS*, ne nécessitent pas d'adaptation supplémentaire car elles reposent explicitement sur le cumul des informations individuelles [44, 53, 87, 132]. La différence totale est donc évaluée indépendamment de la variabilité individuelle de la différence. Au sein des méthodes globales, *GlobalTest* est appropriée pour l'étude de groupes apportant une réponse variable, car la procédure

repose sur l'évaluation de la variabilité de la réponse individuelle. Par contre, *GlobalTest* n'est pas prévu pour détecter des groupes uni-directionnels corrélés, car la variabilité de la réponse individuelle y est nulle (question Q_U^C : « quels sont les groupes uni-directionnels dont les membres présentent une réponse corrélée ? ») [63].

En conclusion à cette comparaison théorique des méthodes actuelles, il apparaît donc que seules les procédures en deux étapes, auxquelles peut être assimilée *SAMGS*, sont capables de répondre simultanément à ces 3 critères et de répondre le plus complètement possible à la question Q_0. Cependant, seules certaines combinaisons entre la statistique individuelle et la statistique de groupe le permettent, mais les logiciels développés ne proposent pas systématiquement ce choix. De plus, la prise en compte de la direction n'est pas équivalente entre les méthodes présentées (la statistique *maxmean*, par exemple, tient uniquement compte du sous-groupe directionnel qui présente la plus forte réponse)[44, 53, 87, 132]. Les méthodes globales, quant à elles, ignorent totalement l'effet du niveau d'expression et sont donc susceptibles de favoriser les groupes dont les membres impliqués sont fortement exprimés. Seule la procédure *SAMGS* répond à l'ensemble des critères mentionnés [44], mais celle-ci implique une étape supplémentaire de stabilisation de la variance sur base de la procédure individuelle de la méthode *SAM*, dont les performances sont variables (voir Résultats, première partie), et qui nécessite l'ensemble des gènes pour être évaluée [92, 136, 152]. En complément, les procédures compétitives s'avèrent inappropriées pour répondre à la question Q_0, car elles ignorent les groupes qui présentent une différence si celle-ci est comparable pour plusieurs groupes [53, 80, 87, 106, 114, 131, 132].

Deux approches peuvent être envisagées pour répondre complètement à la question Q_0. D'une part, nous pouvons utiliser plusieurs méthodologies qui répondent à des questions complémentaires. A titre d'exemple il est possible d'étudier séparément les groupes au sein desquels la réponse est variable et les groupes uni-directionnels. Comme perspective d'extension de la méthode *GlobalAncova*, Mansmann & Meister proposent d'étudier l'interaction entre l'effet de la condition et la composition du groupe, pour prendre en compte la variabilité de la réponse individuelle [104]. Nous pensons cependant que l'hypothèse associée à un tel test est également incomplète, car elle ignore les groupes uni-directionnels pour lesquels la réponse des membres est corrélée (interaction nulle).

Nous présenterons une seconde démarche, qui vise à répondre complètement à question Q_0 sur base d'un seul test auto-suffisant. Pour exploiter au mieux les données d'expression, le modèle envisagé repose sur une stratégie dite globale, car les méthodes en deux étapes

sont caractérisées par une perte d'information liée à la substitution des données par une statistique individuelle. La procédure *ANOVA-2* est adaptée pour l'étude de groupes uni-directionnels. Au cours des prochains paragraphes, nous présenterons la méthodologie *FAERI*, développée par similarité avec l'*ANOVA-2*. Le principe à l'origine de la méthode *FAERI* est de cumuler les hypothèses envisagées en une seule statistique représentative (F^*), capable de détecter tous les types de groupes. La démarche suivie vise donc à cumuler les avantages des modèles envisagés par *GlobalTest* (groupes variables), *GlobalAncova* (groupes uni-directionnels) et *SAMGS* (indépendance vis-à-vis du niveau d'expression), en se libérant de leurs limitations respectives.

IV.B.3. La méthode ANOVA-2

L'analyse de la variance, ou *ANOVA*, repose sur l'étude de la variabilité en relation avec un ou plusieurs critères, pour déterminer si ce(s) critère(s) sont responsables des mesures observées. Le modèle le plus simple, reposant sur un seul critère, est équivalent au test du *t* de STUDENT [130]. En ce qui concerne l'analyse de l'expression différentielle, une *ANOVA* à un critère permet donc d'évaluer l'implication des gènes entre deux ou plusieurs conditions. L'*ANOVA*, considérée avec deux critères de classification, permet d'envisager l'étude de l'expression différentielle de groupes de gènes, le premier critère étant défini par les gènes membres du groupe, et le second critère étant défini par la condition expérimentale testée.

Le modèle que nous avons choisi pour tester la validité de cette approche considère que les données d'expression observées peuvent être expliquées par une dépendance des mesures d'intensité vis-à-vis du gène, vis-à-vis de la condition, et en tenant compte également d'une inter-dépendance entre ces deux paramètres, car chaque gène est susceptible d'offrir une réponse différente à la condition testée.

La représentation mathématique de ce modèle est formulée dans l'équation IV.B.1.

$$X_{(ij)k} = \mu + a_i + b_j + ab_{ij} + E_{(ij)k} \quad \text{(Equ. IV.B.1)}$$

où a symbolise l'effet du gène, b symbolise l'effet de la condition, ab symbolise l'effet de leur interaction, et E représente les effets résiduels non modélisés.

La comparaison de ces différents critères repose sur l'expression du rapport entre les carrés moyens associés à chaque critère, et la statistique obtenue est évaluée par rapport à la distribution de la statistique F de FISHER associée aux degrés de libertés appropriés (Equations IV.B.2 à IV.B.4).

$$F_a = \frac{MS_a}{MS_E} \sim F\left(n_a - 1 ; n_a n_b \left(n_{repl} - 1\right)\right) \quad \text{(Equ. IV.B.2)}$$

$$F_b = \frac{MS_b}{MS_E} \sim F\left(n_b - 1 ; n_a n_b \left(n_{repl} - 1\right)\right) \quad \text{(Equ. IV.B.3)}$$

$$F_{ab} = \frac{MS_{ab}}{MS_E} \sim F\left(\left(n_a - 1\right)\left(n_b - 1\right) ; n_a n_b \left(n_{repl} - 1\right)\right) \quad \text{(Equ. IV.B.4)}$$

où n_a est le nombre de gènes, n_b est le nombre de conditions, et n_{repl} est le nombre de

mesures effectuées pour chaque gène dans chaque condition (le nombre de réplicats). *MS* symbolise le carré moyen (*Mean Square*) et est calculé conformément à la procédure décrite dans la section VI.B.3.a. page 285 de ce travail, grâce à un algorithme optimisé.

Il est important de mentionner également que lorsque l'effet de l'interaction est significatif, aucune décision ne peut être prise sur les autres critères. L'étude de l'effet associé à la condition rejoint les méthodologies *GlobalTest* et *GlobalAncova* [63, 104]. L'étude de l'interaction rejoint les perspectives émises par MANSMANN & MEISTER (*GlobalAncova*), qui considèrent que les groupes au sein desquels les gènes présentent une réponse différente sont biologiquement intéressants [104]. Selon nous, ce test doit être accompagné d'un test sur la condition, car il est impossible de détecter des changements uni-directionnels corrélés en étudiant uniquement l'interaction. Conscients de cette limitation, nous présentons dans le prochain paragraphe la stratégie d'analyse de la méthode *FAERI*, conçue pour répondre simultanément à tous les cas de figures possibles avec un seul test.

IV.B.4. La méthode FAERI

IV.B.4.a. Introduction

L'analyse de groupes de gènes est influencée par la direction de la réponse individuelle des membres et par la présence éventuelle d'une corrélation entre les gènes [21, 63, 64, 98, 104]. En plus de ces deux critères, nous souhaitons en considérer un troisième, le niveau d'expression individuel, pour éviter de donner un poids trop important aux gènes les plus exprimés. Pour répondre à ce critère, nous proposons une étape additionnelle de préparation des données, dont l'objectif est de se libérer de l'effet du niveau d'expression. Pour répondre au critère directionnel, nous proposons de transformer les données pour n'en conserver que la composante absolue de la réponse, et caractériser globalement la force de la réponse du groupe indépendamment de sa structure.

IV.B.4.b. Le niveau d'expression

Les différents gènes d'un même groupe peuvent être exprimés à des niveaux d'expressions très différents. D'un point de vue biologique, l'implication d'un groupe de gènes faiblement exprimés peut avoir un impact plus important sur la physiologie qu'un groupe de gènes fortement exprimés ! Il importe donc de se concentrer sur les informations liées au différences de réponse entre les deux conditions, quel que soit le niveau d'expression des différents gènes. Dans la perspective de l'analyse de groupes de gènes, nous proposons de standardiser les données pour que tous les gènes aient *a priori* le même potentiel de réponse entre les deux conditions comparées.

Cette standardisation des données peut être effectuée à l'aide d'une procédure classique de réduction des données en valeurs Z, souvent utilisée en statistique. La procédure mathématique utilisée repose sur l'équation IV.B.5, pour chaque gène, sur base de la moyenne et de la variance observées sur l'ensemble des données qui le concernent. Les données réduites sont distribuées autour d'une valeur moyenne égale à 0 avec une variance ramenée à 1. L'information relative à la différence entre les conditions, pour chaque gène, est conservée. Le résultat d'une analyse individuelle, sur base du t de STUDENT (cas particulier de l'*ANOVA* avec un seul critère de classification), génère strictement les mêmes résultats.

$$Z = \frac{X - M_X}{S_X} \quad \text{(Equ. IV.B.5)}$$

Les résultats obtenus sur des données réelles au terme de cette opération révèlent qu'un grand nombre de groupes de gènes paraissent significatifs soit pour l'effet de la condition, soit pour l'interaction, soit pour les deux (non représenté). Il apparaît évident que la caractérisation de l'implication d'un groupe est partagée entre ces deux indicateurs. Au sein du prochain paragraphe, nous présentons la stratégie suivie pour cumuler cette information au sein d'un seul indicateur, représentatif de la réponse du groupe, grâce à la réduction directionnelle.

IV.B.4.c. Direction de la réponse

Chaque gène étant régulé par plusieurs facteurs, plusieurs gènes peuvent être activés ou inhibés ensembles dans un contexte donné, mais ils peuvent aussi apporter une réponse variable selon le contexte biologique étudié. Il n'est donc pas toujours possible d'identifier la direction réelle des différences pour chaque gène. L'hypothèse nulle à l'origine de l'analyse *ANOVA-2* se formule par « le niveau moyen d'expression est-il différent entre les deux conditions comparées ? » ($\mathcal{Q}_U$). Notre objectif est de répondre à la question « L'ensemble des différences individuelles, quelles que soient leurs directions, est-elle significative lorsque l'on compare les deux conditions étudiées ? » ($\mathcal{Q}_B$).

Au terme de la première étape décrite ci-dessus, l'interaction observée sur les données réduites s'explique par la variabilité de la réponse des différents gènes. MANSMANN & MEISTER, en perspective d'extension de la méthode *GlobalAncova*, proposent d'évaluer la question $\mathcal{Q}_B$ par l'étude de l'interaction (la variabilité de la réponse), que nous pouvons symboliser par $\mathcal{Q}_I$. Ils considèrent ainsi pouvoir détecter les groupes bi-directionnels, et fournir un résultat plus complet [104]. Le test envisagé serait toutefois limité à l'étude de la variabilité de la réponse, et inapproprié pour la détection de groupes dont les membres sont corrélés (que nous noterons $\mathcal{Q}^{cor}$).

Deux approches peuvent être envisagées pour répondre complètement à la question $\mathcal{Q}_B$: étudier séparément les gènes dont la direction de la différence est la même, ou alors étudier la différence absolue, quelle qu'en soit la direction. La première solution consiste à définir des groupes de gènes « activés » ou « inhibés » pour pouvoir procéder à une analyse

correcte. Cette démarche est tributaire des connaissances actuelles, incomplètes, mais est utile à des fins diagnostiques, afin de déterminer si un patient présente une pathologie en étudiant la réponse d'un groupe appelé « signature ». La seconde solution est abordée par plusieurs méthodes via la définition d'une statistique individuelle indépendante de la direction [44, 53, 114, 132].

Dans le cadre d'une procédure multivariée, nous proposons de répondre à la question Q_B en étudiant la réponse absolue des membres, grâce à une réduction directionnelle des données. Cette opération s'effectue en multipliant par -1 toutes les données individuelles relatives à une sous-expression (Equation IV.B.6). N'ayant pas de connaissances biologiques suffisamment complètes à notre disposition à ce jour, la correction proposée doit être réalisée de manière empirique, au départ des données d'expression étudiées.

$$X_i^D = sign\left(D_i\right) X_i \quad \text{(Equ. IV.B.6)}$$

où X_i dénote les valeurs d'expressions associées au *probeset* i, et D_i est la différence des moyennes des mesures d'expression réalisées dans les deux conditions comparées, pour le *probeset* i.

Pour analyser le jeu de données corrigé, *FAERI* évalue ensuite une statistique F sur le même principe que l'*ANOVA-2*. La transformation des données, pour n'en conserver qu'une seule composante différentielle, indépendante de sa direction, présente plusieurs conséquences sur la valeur de la statistique F évaluée. L'échantillonnage aléatoire d'un nombre réduit de valeurs individuelles génère un « biais » conduisant à mesurer une différence individuelle. En moyenne, ce biais est équiprobable dans les deux directions, avec pour conséquence, lors d'une analyse *ANOVA-2*, d'un effet moyen nul. Par contre, à l'issue de la transformation directionnelle, l'effet global de ce biais est cumulatif, et s'ajoute à la différence mesurée. La seconde conséquence de cette transformation dérive de ce biais : puisqu'il est cumulatif, la statistique F évaluée est dépendante de la taille des groupes de gènes, et est d'autant plus grande que le groupe comporte de membres.

La figure IV.B.1 illustre la dépendance entre la statistique F calculée et le nombre de membres, respectivement sur des données générées aléatoirement (partie supérieure) et sur le jeu de données biologiques E-GEOD-7429 (partie inférieure). Les statistiques calculées sur base de données biologiques sont plus élevées, particulièrement suite à la réduction Z. Cette étape influe donc fortement sur les résultats de l'analyse (la variabilité résiduelle est beaucoup plus petite à l'issue de cette étape). La réduction directionnelle s'accompagne d'une dépendance avec le nombre de nombre.

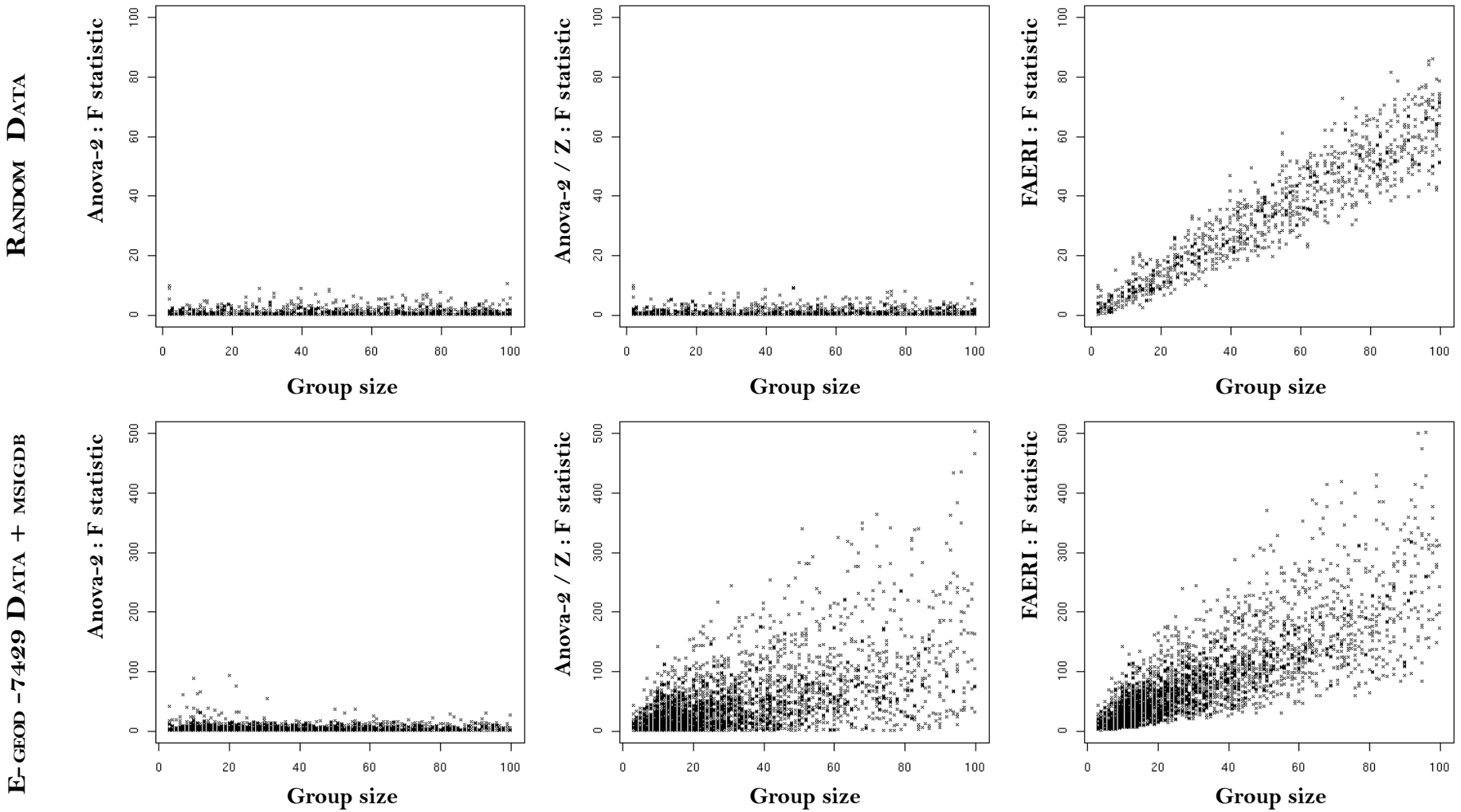


Figure IV.B.1 : Illustration de la distribution de la statistique F évaluée par comparaison avec la taille des groupes, en utilisant la procédure ANOVA-2, sur les données d'expression initiales (gauche), sur les données standardisées (centre), et sur les données standardisées et rendues uni-directionnelles (droite, procédure FAERI). Les graphiques de la région supérieure ont été générés au départ de données aléatoires et montrent que l'étape de réduction directionnelle induit une dépendance vis à vis du nombre de membres. Les graphiques de la région inférieure illustrent les résultats obtenus sur des données réelles (jeux de données E-GEOD-7429), et montrent d'une part l'impact de la standardisation des données propres à chaque gène, et d'autre part la dépendance vis-à-vis du nombre de membres suite à la réduction directionnelle.

L'hypothèse à l'origine de l'analyse ne correspond pas à l'hypothèse utilisée classiquement en analyse de la variance (distribution normale des données), si bien que la significativité du résultat ne peut être évaluée sur base de la distribution théorique de la statistique F de FISHER, utilisée en *ANOVA*. Pour distinguer la statistique évaluée par la procédure *FAERI*, nous désignerons par F^* la statistique évaluée. Pour estimer objectivement sa significativité, une distribution de valeurs F^* propres à notre stratégie d'analyse doit être évaluée, en tenant compte du nombre de gènes présents au sein du groupe.

A l'issue de ces deux étapes de réduction des données, la statistique F^* employée est représentative des différences recherchées sur base d'un raisonnement biologique. Chaque gène apporte une information sur son expression différentielle, indépendamment du sens de cette différence. Au sein du prochain paragraphe, nous présenterons deux procédures différentes d'évaluation de la distribution nulle de la statistique F^*, sur base de données aléatoires, ou sur base de permutations des échantillons.

IV.B.4.d. Evaluation de la significativité

Au cours de la mise au point de la procédure *FAERI*, par homologie avec l'*ANOVA*, les particularités des jeux de données biologiques ont été prises en considération, en envisageant les différents cas de figure possibles. Les adaptations apportées à la procédure (réduction en valeurs Z et réduction directionnelle) conduisent à une statistique F^* distribuée différemment de la statistique classique de FISHER, et dépendante de la taille du groupe. Il est donc nécessaire de déterminer la distribution nulle de la statistique F^* de *FAERI*. Les deux solutions couramment envisagées par les statisticiens reposent sur l'usage de permutations, d'un part, et de données aléatoires, d'autre part. Ces deux approches ont été suivies pour évaluer la significativité de la méthode *FAERI*, et sont rapportées ci-dessous.

Nous souhaitons savoir, pour chaque groupe de gènes, si les mesures effectuées sur ses membres diffèrent entre deux profils d'expressions comparés. Dans la procédure *FAERI*, il s'agit d'un test « auto-suffisant » puisque seules les données relatives au groupe sont utilisées. Dans pareil cas, le modèle de permutations approprié repose sur un ré-échantillonnage des labels associés aux conditions comparées [64, 132]. Etant donné que l'hypothèse de départ est qu'aucun groupe n'est impliqué entre les conditions comparées,

que ses membres soient corrélés ou non, les permutations sont réalisées indépendamment pour chaque gène.

Les jeux de données issus de puces à ADN étant coûteux, le nombre de mesures réalisées est souvent réduit (quelques réplicats), et le nombre de permutations disponibles est par conséquent limité. Cette limitation affecte la résolution des *p-values* obtenues, qui suit une distribution discrète.

L'étape de réduction des données en valeurs Z ouvre une autre perspective relative à l'évaluation de la distribution de référence de la statistique F^* de *FAERI*. En effet, puisque l'hypothèse nulle de la méthode est qu'il n'y a pas de différence entre les deux conditions, et puisque nous postulons que les données d'expression individuelles suivent une distribution normale, la standardisation des données permet d'établir la distribution de référence au départ de données générées aléatoirement sur base d'une distribution normale. En analysant ces données aléatoires avec *FAERI*, le même processus de réduction Z et de réduction directionnelle s'y appliquent. Dans le respect de l'hypothèse nulle, les données réduites sont équivalentes et la statistique F^* suit la même distribution. Dans le cas contraire, pour un groupe de gènes présentant une différence entre les conditions comparées, la distribution de la statistique F^* présente des valeurs plus élevée que celles dues au hasard.

Cette opportunité de calculer une distribution de référence au départ de données aléatoires libère la méthode *FAERI* du nombre limité de permutations accessibles. De plus, pour chaque stratégie expérimentale (nombre de réplicats) et chaque taille de groupe de gène, la distribution de référence peut être calculée une fois pour toutes, contrairement aux permutations qui doivent être réalisées pour chaque groupe et pour chaque jeu de données. Enfin, la seule limitation restante du point de vue du niveau de discrétisation des *p-values* obtenues repose uniquement sur le temps de calcul nécessaire pour analyser les données aléatoires.

Tenant compte du biais systématique introduit au cours de la réduction directionnelle réalisée par la méthode *FAERI*, les deux procédures d'évaluation envisagées permettent de se libérer de la dépendance observée entre la statistique F^* évaluée et la taille du groupe de gènes, et d'attribuer une *p-value* à chaque groupe. La figure IV.B.2 montre que la *p-value* évaluée est effectivement indépendante du nombre de gènes membres, sur base de

permutations, ou sur base de données aléatoires (jeu de données biologique).

La figure IV.B.3 montre également que les *p-values* obtenues sur base des deux modèles sont parfaitement corrélées dans le cas de données simulées, et sont différentes sur un jeu de données biologiques.

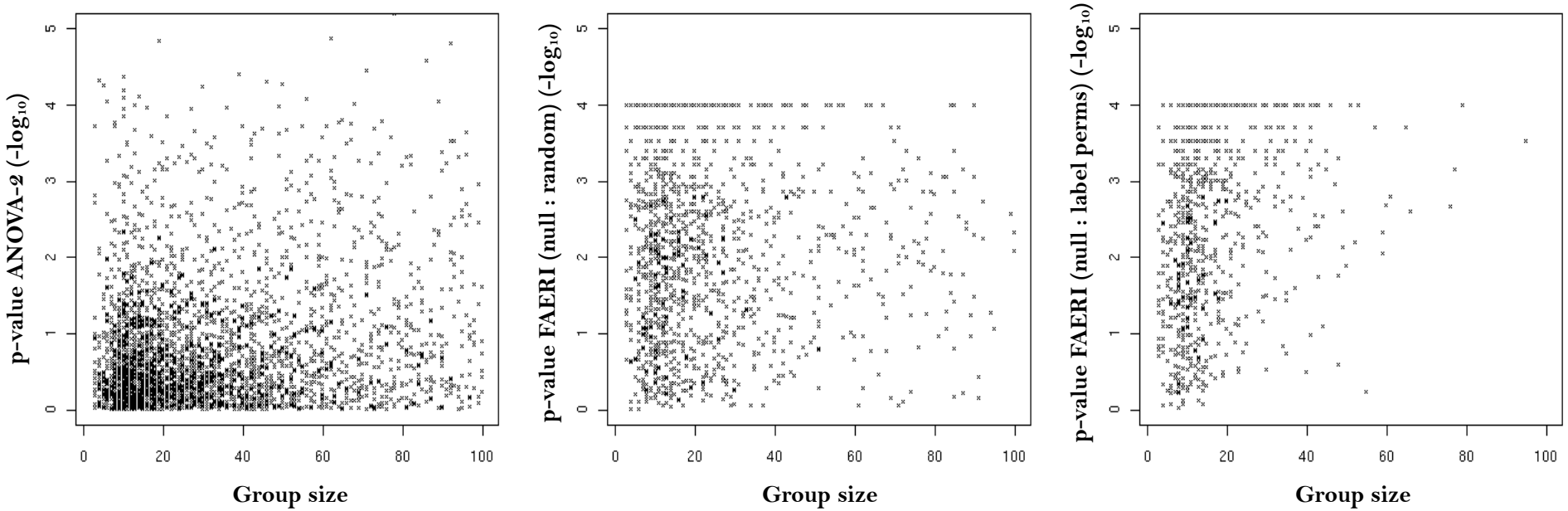


Figure IV.B.2 : Illustration du logarithme des p-values évaluées par l'*ANOVA-2* (gauche), par *FAERI* sur base de données aléatoires (centre) ou de permutations (droite), en fonction du nombre de membres (Jeu de données réelles E-GEOD-7429). Les graphiques présentés au centre et à droite montrent que les deux procédures envisagées pour évaluer la significativité du test *FAERI* fournissent des p-values indépendantes de la taille des groupes.

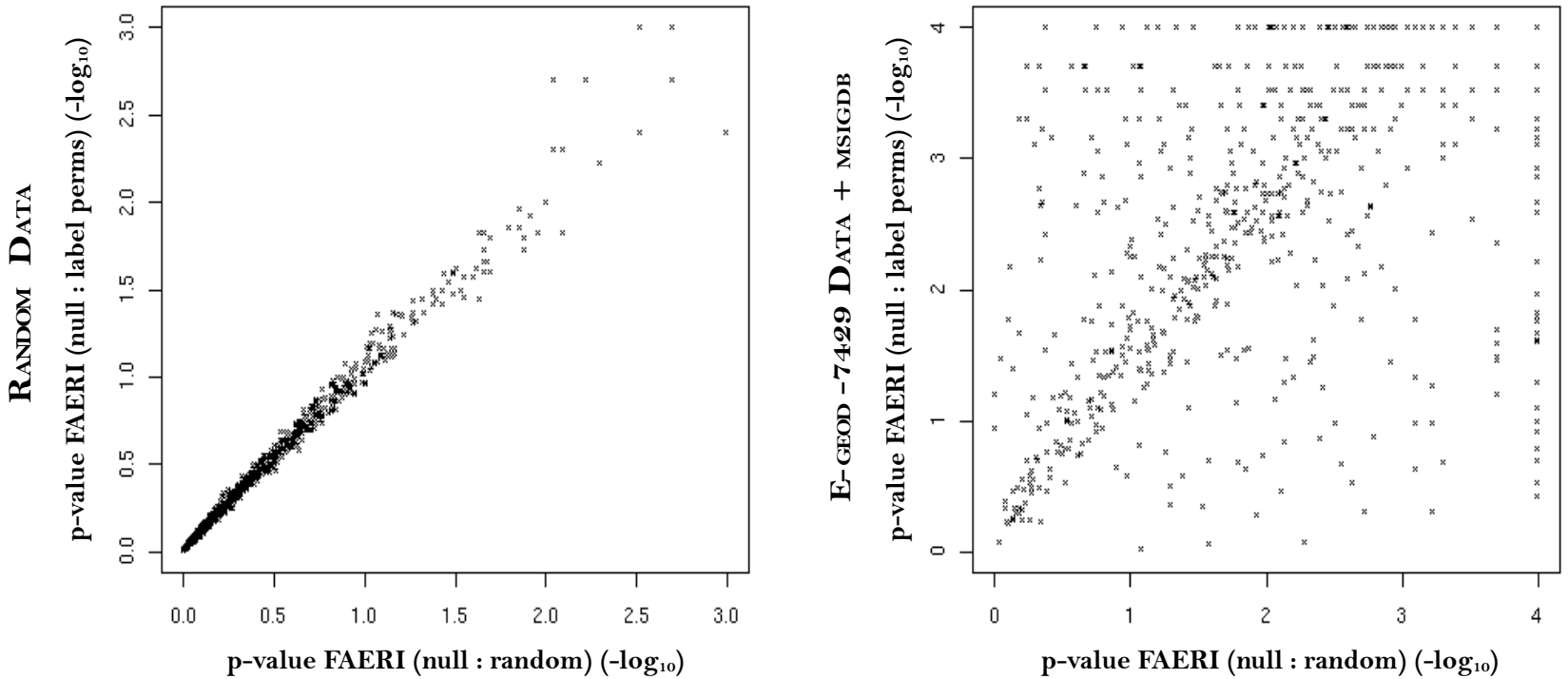


Figure IV.B.3 : Comparaison du logarithme des p-values évaluées par FAERI sur base de données aléatoires ou de permutations. Le graphique de gauche illustre la comparaison des p-values obtenues lors de l'analyse de données simulées. Le graphique de droite illustre les résultats obtenus lors de l'analyse de données réelles (E-GEOD-7429), et montre que la distribution nulle évaluée par les deux procédures est différente dans le cas de données réelles, mais une partie des groupes présentent néanmoins une p-value similaire (diagonale).

IV.B.5. Evaluation des performances

Afin de valider la démarche suivie au sein de la méthode *FAERI*, et de faire le point sur les performances des différentes méthodes disponibles actuellement, plusieurs stratégies d'évaluation sont présentées dans les prochains paragraphes.

D'une part, les performances ont été évaluées sur base de données aléatoires indépendantes. Des gènes y sont simulés avec ou sans expression différentielle, et des groupes de gènes y sont aléatoirement définis parmi ces deux catégories de gènes. D'autre part, des simulations sont menées plus avant, sur base de distributions normales multivariées, pour étudier l'effet de la corrélation entre les gènes d'un même groupe sur les performances du test. Après avoir présenté les résultats de ces simulations, les résultats des différentes méthodes seront discutés sur base de trois jeux de données relatifs à l'hypoxie, avec plusieurs sources de définitions de groupes de gènes. La pertinence qualitative des résultats y sera présentée, et des mesures quantitatives de corrélation entre les différents jeux de données serviront à caractériser les performances des méthodes de groupes.

Les méthodes sélectionnées pour ces études comparatives sont représentatives des différentes approches envisagées en analyse de groupe. *GSEA*, méthode pionnière, répandue et critiquée dans la littérature, a été retenue comme représentante des méthodes basées sur un score d'enrichissement évalué en parcourant la liste totale des gènes [106, 131]. Les approches disponibles dans le package *GSA* sont utilisées pour comparer les statistiques *mean*, *absmean* et *maxmean* [53]. La méthode *GlobalTest* [63], qui utilise les données d'expression brutes sur base d'un modèle multivarié, a également été sélectionnée, de même que la méthode *SAMGS* [44], représentante d'une approche multivariée similaire au test du T^2 de Hotelling [70, 90]. La procédure *ANOVA-2* et la méthode *FAERI* ont été évaluées par comparaison avec ces méthodes. Lorsque plusieurs modèles de permutations sont disponibles pour l'attribution de *p-values*, seuls les modèles non compétitifs sont retenus, pour que les hypothèses testées soient équivalentes.

IV.B.5.a. Simulations de données aléatoires indépendantes

Le premier scénario de simulation envisagé, sur base d'une distribution normale, correspond à l'analyse de groupes définis aléatoirement parmi un ensemble de gènes simulés différemment entre deux conditions. Dix niveaux de différences individuelles ont été simulés pour ces gènes, au nombre de 400 au total (200 activés et 200 réprimés). Pour la composition de l'hypothèse nulle, 19.600 gènes ont été simulés. Mille gènes ont été simulés sans différence d'expression, pour dix niveaux d'expression différents (de 1 à 10), et 9600 gènes non exprimés ont été simulés. La variance de tous les gènes est simulée sur base d'une distribution gamma avec une moyenne égale à 2 et une variance égale à 2. Le jeu de données simulées comporte donc 20.000 gènes, dont 2% est différentiellement exprimé, avec des niveaux d'expression et des différences d'expression variables. Pour générer les groupes de gènes, le scénario envisagé repose sur une analyse de 500 groupes, reproduite 100 fois, dont 25 groupes sont différentiellement exprimés. Ceux-ci sont générés aléatoirement parmi les 400 gènes simulés différentiellement. Les 475 autres groupes sont générés au départ des 19.600 gènes qui ne sont pas différentiellement simulés. Au total, les analyses réalisées concernent donc 50.000 groupes, dont 2.500 sont simulés différentiellement (5%), avec des membres simulés à plusieurs niveaux et dont la différence simulée est variable.

La figure IV.B.4 présente les performances obtenues par les méthodes actuellement disponibles, en définissant aléatoirement des groupes de « gènes impliqués » parmi l'ensemble des données simulées différemment entre les deux conditions, et des groupes « contrôles » construits aléatoirement parmi les données simulées sans différence entre les deux conditions. Les groupes définis sont de taille fixe (A) ou aléatoire (B), en simulant des données unidirectionnelles (A et B) ou bidirectionnelles (C). La comparaison des graphiques A et B illustre l'impact de la taille des groupes, la comparaison des graphiques B et C illustre l'impact de la bi-directionnalité des données. *GSEA* et *SAMGS* ne sont pas représentées car les scripts fournis par leurs auteurs provoquent des erreurs à l'exécution.

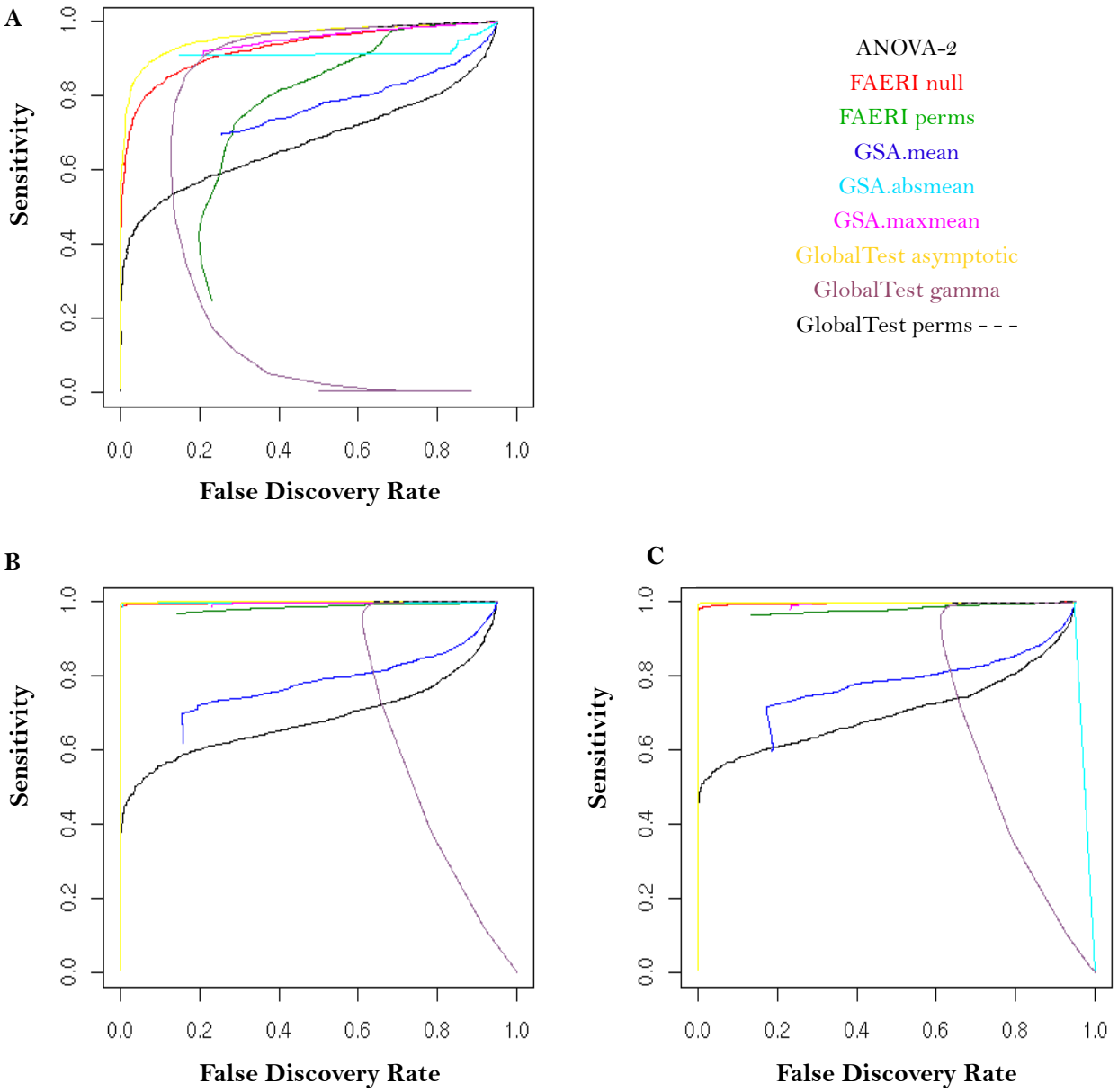


Figure IV.B.4 : Comparaison des performances des méthodes d'analyse de groupes actuellement disponibles. Chaque graphique compare la sensibilité au FDR, et illustre la capacité des méthodes à découvrir la vérité en fonction du prix à payer pour la découvrir. Les méthodes les plus performantes sont associées aux courbes les plus proches du coin supérieur gauche. En haut à gauche: les groupes définis sont uni-directionnels de taille constante, égale à 3 ; en bas: les groupes définis sont uni-directionnels et de taille aléatoire (à gauche) ou bi-directionnels et de taille aléatoire (à droite). Globalement, les méthodes FAERI null et GlobalTest asymptotic sont les plus performantes. Les performances de GlobalTest évalué par des permutations sont incapable de discriminer les groupes avec un taux d'erreur inférieur à 60% (FDR). Pour le même nombre de permutations, FAERI.perms est capable de détecter des groupes jusqu'à un taux d'erreur inférieur 20%. Lors de l'analyse de groupes de tailles aléatoire, GlobalTest.asymptotic, FAERI.null, FAERI.perms, GSA.absmean et GSA.maxmean forment le groupe de tête. Lors de l'analyse de groupes bi-directionnels de taille aléatoire, GSA.absmean est défavorisée.

L'examen de la figure IV.B.4 fourni les informations suivantes :

- ☞ les méthode *FAERI.null* et *GlobalTest.Asymptotic* sont les plus performantes, dans tous les cas ;
- ☞ lors de l'analyse de groupes uni-directionnels de tailles aléatoire, les meilleurs méthodes sont *GlobalTest.asymptotic*, *FAERI.null*, *FAERI.perms*, *GSA.absmean* et *GSA.maxmean ;*
- ☞ *GSA.absmean* quitte le groupe de tête lorsque les groupes générés sont bi-directionnels ;
- ☞ les performances de *GlobalTest*, évaluées par des permutations montrent que la méthode est incapable de discriminer les groupes avec un taux d'erreur inférieur à 60% (*FDR*). Pour le même nombre de permutations, *FAERI.perms* est capable de détecter des groupes avec un taux d'erreur inférieur 20% ;
- ☞ les performances de *FAERI.perms* sont inférieures aux performances de *FAERI.null* lorsque les groupes générés sont de petite taille (3 membres).

Les premières observations rapportées correspondent à différents scénarios dont le seul point commun est l'absence de corrélation entre les membres des groupes, car ceux-ci sont générés indépendamment. *GlobalTest* repose sur un modèle qui favorise la détection de groupes au sein desquels les membres offrent une réponse variable. Les tests réalisés correspondent donc à l'hypothèse envisagée par les auteurs de cette méthode. La méthode *FAERI* a été développée pour analyser les groupes de différentes natures, et apparaît appropriée pour l'analyse de groupes dont les membres ne sont pas corrélés, que les données soient uni-directionnelles ou bi-directionnelles, peu importe la taille des groupes, et peu importe le niveau d'expression individuel des gènes simulés.

Pour compléter l'évaluation des performances des méthodes en regard de la question Q_0, nous présenterons, dans le prochain paragraphe, un tableau comparatif des performances associées à chaque méthode, en tenant compte de la direction de la réponse, de la proportion de membres différentiellement exprimés, et de la présence d'une corrélation entre les membres. Cette étude plus détaillée des performances permettra de mieux discriminer les méthodes, et de démontrer la pertinence de la démarche suivie lors du développement de la méthode *FAERI* pour chaque critère envisagé.

IV.B.5.b. Simulations de données aléatoires corrélées

Plusieurs auteurs rapportent dans leurs publications la nécessité d'étudier le comportement des différentes méthodes lorsque les données d'un groupe sont corrélées [53, 64, 98]. A cette fin, et dans le but de décrire les résultats obtenus par différentes méthodes, MARIT ACKERMANN a mis au point, en 2009, une stratégie d'analyse des performances sur base de 9 groupes de « gènes »[21]. Les données relatives à chacun de ces groupes sont générées aléatoirement à l'aide d'une distribution normale multivariée, et de la définition d'une matrice de corrélation. Les différents scénarios envisagés sont présentés au sein de la table IV.B.2. Dans chaque cas, 10 réplicats sont simulés pour chaque condition, et le groupe comporte 20 membres [21].

	Différence d'expression	Corrélation	Diff. Exprimés	Sur-exprimés	Sous-exprimés	Design
Set n°1	0.75	0.6	20	20	0	uni
Set n°2	0.75	0	20	20	0	uni
Set n°3	0	0	0	0	0	
Set n°4	0.75	0.6	10	10	0	uni
Set n°5	0.75	0	10	10	0	uni
Set n°6	1	0.6	20	10	10	bidir
Set n°7	1	0	20	10	10	bidir
Set n°8	1	0.6	10	5	5	bidir
Set n°9	1	0	10	5	5	bidir

Table IV.B.2
Définition des séries de mesures utilisées pour évaluer les performances des méthodes d'analyse de l'expression différentielle de groupes de gènes.

Les comparatifs de performances publiés par ACKERMANN s'attachent toutefois à la comparaison des choix réalisés aux différentes étapes de l'analyse, en utilisant un seuil de sélection de 0.05, et ne fourni pas d'évaluation comparative des performances des méthodes existantes [21].

Nous avons reproduit la simulation proposée par ACKERMANN, et avons analysé les performances des différentes méthodes actuellement disponibles, ainsi que de l'*ANOVA-2*, et de la méthode *FAERI*. La table IV.B.3 présente les résultats obtenus en suivant le scénario uni-directionnel, lorsque le seuil de sélection sur les *p-values* est fixé à 0.1, 0.01, et 0.001. Pour chaque groupe de gènes simulé, la simulation a été reproduite 100 fois. Les valeurs présentées fournissent le nombre de détections du groupe parmi ces 100 simulations.

La comparaison des résultats obtenus illustre l'effet de la corrélation entre les membres d'un groupe sur les performances de l'analyse. Pour chaque méthode envisagée, la présence d'une corrélation intra-groupe réduit les performances de l'analyse. Pour les sets les plus simples, uni-directionnels, la méthode *ANOVA-2* fourni les meilleurs résultats, suivie par la méthode *FAERI* (sets 1, 2, 4), ou *GlobalTest* (set 5 – gamma ou permutations).

De plus, le modèle de simulation le plus simple, représenté par les groupes n°1, 2 et 3 montre la supériorité des méthodes *ANOVA-2* et *FAERI* sur toutes les autres méthodes étudiées. En effet, *FAERI* détecte le set n°1 dans 62% des cas, avec un seuil de 0.001, sans faux positifs (0% pour le set n°3). A ce seuil, *FAERI* découvre environ 5 fois plus de groupes que la meilleure méthode publiée (*SAMGS* avec *q-value*: 13%). A titre de comparaison, la détection de 60% du set n°1 par *SAMGS* et *GlobalTest* implique un seuil de sélection 100 fois plus élevé. La comparaison des set n°4 et 5 illustre la capacité des méthodes à détecter des groupes dont la moitié des membres ne sont pas différentiellement exprimés. *FAERI* détecte dans ce cas 43% des groupes, contre 15% pour la meilleure méthode publiée : *GlobalTest* lorsque la significativité est calculée sur base d'une distribution gamma. Les méthodes *ANOVA-2* et *FAERI* sont donc les plus appropriées pour l'analyse de groupes de gènes uni-directionnels.

La table IV.B.4 illustre les résultats obtenus lorsque les données simulées impliquent des différences d'expression dans les deux directions (sets 6, 7, 8 et 9). Le comportement des méthodes diffère : les méthodes *ANOVA-2*, *GSA* utilisant la moyenne comme statistique de groupe (*GSA.mean*), ainsi que *GSEA*, s'avèrent inadaptées. Dans tous les cas, la méthode *FAERI* génère les meilleurs résultats, à l'exception du set n°9 qui est mieux découvert par *SAMGS* pour un seuil inférieur à 0.05. La présence d'une corrélation entre les membres réduit les performances, ainsi qu'observé pour les autres sets.

Pour simplifier l'interprétation des performances des méthodes en regard de la nature des groupes analysés, la table IV.B.5 présente le classement des méthodes pour chaque catégorie de groupes définie, pour un seuil de sélection placé à 0.01. La statistique utilisée est l'*accuracy*, défini par le rapport entre d'une part le nombre de vrais positifs et vrais négatifs, et d'autre part le nombre de groupes testés. Les scores supérieurs à 75% sont surlignés en gris. Les performances ont été évaluées sur base des tables IV.B.3 et IV.B.4.

		Cut-off		a2.fixed	faeri.fixed.null	faeri.fixed.perms	GSA.mean.*	GSA.absmean.*	GSA.maxmean.*	globaltest.asymptotic	globaltest.gamma	globaltest.permutations	gsea.pval	gsea.fdr	samgs.pval	samgs.qval
H_0		**0.001**	**Set3**	0	0	0	0	0	0	0	0	0	0	0	0	1
		0.01	**Set3**	1	0	0	0	0	0	0	0	0	0	0	0	2
		0.1	**Set3**	3	3	3	12	0	0	2	5	3	2	1	2	27
Uni	100% DE	**0.001**	**Set2**	100	96	97	15	25	12	54	96	95	69	20	91	94
		0.01	**Set2**	100	99	99	15	25	12	95	99	97	99	77	98	99
		0.1	**Set2**	100	100	100	78	72	64	99	99	99	100	100	99	100
	50% DE	**0.001**	**Set5**	72	40	36	0	2	0	3	64	49	14	2	39	56
		0.01	**Set5**	89	75	78	0	2	0	42	89	84	42	25	81	84
		0.1	**Set5**	100	94	94	7	17	3	93	99	98	93	84	96	99
Uni + Cor	100% DE	**0.001**	**Set1**	86	62	60	1	2	2	1	10	8	2	0	7	13
		0.01	**Set1**	89	69	70	1	2	2	20	35	25	16	3	23	41
		0.1	**Set1**	95	80	80	11	23	11	58	59	58	46	47	56	73
	50% DE	**0.001**	**Set4**	55	43	39	0	0	0	0	15	7	4	3	4	11
		0.01	**Set4**	65	52	54	0	0	0	18	30	24	13	14	23	43
		0.1	**Set4**	81	65	65	1	13	1	56	60	57	43	47	56	73

Table IV.B.3

Comparaison des performances des méthodes d'analyse de l'expression différentielle de groupes de gènes, sur base du modèle de simulation proposé par M. Ackermann. Chaque set de mesures représenté a été généré 100 fois. Les mesures reprises dans le tableau indiquent, pour chaque méthode, le nombre de détections des différents sets de mesures définis. Tous les groupes générés sont uni-directionnels. Les données relatives à H_0 indiquent le nombre de groupes détectés par hasard (faux positifs).

		Cut-off		a2.fixed	faeri.fixed.null	faeri.fixed.perms	GSA.mean.*	GSA.absmean.*	GSA.maxmean.*	globaltest.asymptotic	globaltest.gamma	globaltest.permutations	gsea.pval	gsea.fdr	samgs.pval	samgs.qval
H_0		**0.001**	**Set3**	0	0	0	0	0	0	0	0	0	0	0	0	1
		0.01	**Set3**	1	0	0	0	0	0	0	0	0	0	0	0	2
		0.1	**Set3**	3	3	3	12	0	0	2	5	3	2	1	2	27
Bi-dir	100% DE	**0.001**	**Set7**	0	100	100	1	67	40	100	100	100	1	0	100	100
		0.01	**Set7**	0	100	100	1	67	40	100	100	100	10	1	100	100
		0.1	**Set7**	6	100	100	22	98	82	100	100	100	69	32	100	100
	50% DE	**0.001**	**Set9**	0	84	82	0	7	2	32	95	93	0	0	86	93
		0.01	**Set9**	0	95	95	0	7	2	93	100	100	1	0	100	100
		0.1	**Set9**	9	100	100	16	48	15	100	100	100	27	8	100	100
Bi-dir + cor	100% DE	**0.001**	**Set6**	0	84	84	0	11	8	5	28	17	6	1	15	25
		0.01	**Set6**	0	88	89	0	11	8	39	50	43	17	8	43	61
		0.1	**Set6**	2	93	93	35	49	54	79	79	78	57	42	78	88
	50% DE	**0.001**	**Set8**	0	63	62	0	2	1	3	30	14	0	0	13	23
		0.01	**Set8**	0	75	76	0	2	1	36	50	42	3	0	40	58
		0.1	**Set8**	3	86	86	28	27	12	75	78	75	31	15	76	85

Table IV.B.4

Comparaison des performances des méthodes d'analyse de l'expression différentielle de groupes de gènes, sur base du modèle de simulation proposé par M. Ackermann. Chaque set de mesures représenté a été généré 100 fois. Les mesures reprises dans le tableau indiquent, pour chaque méthode, le nombre de détections des différents sets de mesures définis. Tous les groupes générés sont bi-directionnels. Les données relatives à H_0 indiquent le nombre de groupes détectés par hasard (faux positifs).

	a2.fixed	faeri.fixed.null	faeri.fixed.perms	GSA.mean.*	GSA.absmean.*	GSA.maxmean.*	globaltest.asymptotic	globaltest.gamma	globaltest.permutations	gsea.pval	gsea.fdr	samgs.pval	samgs.qv
100% uni	**100%**	**99%**	**99%**	57%	62%	56%	**97%**	**99%**	**98%**	**99%**	**88%**	**99%**	**99%**
50 % uni	**94%**	**87%**	**89%**	50%	51%	50%	71%	**94%**	**92%**	71%	62%	**90%**	**92%**
100% uni + cor	**94%**	**84%**	**85%**	50%	51%	51%	60%	67%	62%	58%	51%	61%	70%
50% uni + cor	**82%**	**76%**	**77%**	50%	50%	50%	59%	65%	62%	56%	57%	61%	71%
100% bi-dir	50%	**100%**	**100%**	50%	**83%**	70%	**100%**	**100%**	**100%**	55%	50%	**100%**	**100%**
50 % bi-dir	50%	**97%**	**97%**	50%	53%	51%	**96%**	**100%**	**100%**	50%	50%	**100%**	**100%**
100% bidir + cor	50%	**94%**	**94%**	50%	55%	54%	69%	75%	71%	58%	54%	71%	**80%**
50% bidir + cor	50%	**87%**	**88%**	50%	51%	50%	68%	75%	71%	51%	50%	70%	**79%**
moyenne	71%	**90%**	**91%**	51%	57%	54%	**77%**	**84%**	**82%**	62%	58%	**81%**	**86%**
classement	8	2	1	13	11	12	7	4	5	9	10	6	3

Table IV.B.5 : Caractérisation des performances de chaque méthode, par comparaison du nombre de vrais positifs et vrais négatifs avec le nombre total de test (VP+VN)/(P+N). A titre indicatif, nous en avons également calculé la valeur moyenne sur base des 8 types de groupes étudiés. Le seuil de significativité utilisé pour sélectionner les groupes a été fixé à 0.01. Les valeurs surlignées en gris indiquent les méthodes qui obtiennent un score supérieur à 75%. Les quatre méthodes globales sont les plus appropriées pour l'analyse de groupe, et les méthodes à deux étapes affichent de faibles performances. Parmi les méthodes globales, seules l'ANOVA-2 et FAERI s'avèrent adaptées à l'étude de groupes uni-directionnels corrélés. FAERI et SAMGS s'avèrent les seules méthodes adaptées à l'étude de groupes bi-directionnels corrélés. Globalement, pour l'ensemble des groupes testés, les méthodes qui fournissent les meilleurs résultats sont, dans l'ordre: FAERI (perms), FAERI (null), SAMGS (*q-value*), GlobalTest (gamma), GlobalTest (perms), SAMGS (p-value), GlobalTest (asymptotic). FAERI est la seule méthode adaptée pour l'étude de tous les types de groupes testés.

L'examen de la table IV.B.5 montre que les méthodes qui fournissent les meilleurs résultats, pour chacune des catégories, ont été développées sur base d'un modèle dit « global ». Le classement des méthodes est conforme aux modèles envisagés par ces méthodes :

- ☞ l'*ANOVA-2* s'avère la plus appropriée pour l'étude des groupes uni-directionnels ;
- ☞ *GlobalTest*, qui utilise une statistique représentative de la variabilité de la réponse, est appropriée pour l'étude de groupes uni-directionnels et bi-directionnels non corrélés ;
- ☞ *SAMGS*, en utilisant une statistique individuelle, s'avère inappropriée pour l'analyse de groupes corrélés, car cette information est perdue lors de la procédure suivie.
- ☞ la méthode *FAERI*, développée pour la détection de groupes indépendamment de leur nature, s'avère appropriée pour l'analyse de toutes les catégories testées.

- ☞ les méthodes *ANOVA-2* et *FAERI* apportent une réponse appropriée à l'étude de groupes dont les membres présentent une réponse corrélée, respectivement pour les groupes uni-directionnels et bi-directionnels..
- ☞ les méthodes qui reposent sur une stratégie en deux étapes sont les plus mal classées (*GSA* et *GSEA*).

IV.B.6. Exemple Biologique: cas de l'hypoxie

La validation des performances des méthodes d'analyse de groupe sur des données biologiques réelles est plus complexe que l'étude des performances des méthodes d'analyse individuelles. La définition des groupes sur base de critères biologiques fournit une bibliothèque de groupes dont la structure « mathématique » est variable, en fonction de la taille du groupe, de la corrélation entre les membres (régulation de l'expression des gènes), de la direction de la réponse, de la méthode employée pour définir ces groupes (au départ de la connaissance théorique d'une voie métabolique, au départ de l'observation de signatures de coexpression par clustering ...).

L'évaluation des performances ne peut être réalisée efficacement qu'en étudiant séparément les différents types de groupes, mais il ne nous est pas possible de les « classer » en fonction de leur structure mathématique sans craindre l'introduction d'un biais. Ces évaluations ont donc été réalisées sur base de simulations.

L'analyse d'un jeu de données biologique réel offre la possibilité d'évaluer qualitativement les résultats, et de valider l'utilisation de ces méthodes sur base de leur cohérence biologique avec les connaissances actuelles. Nous avons choisi trois jeux de données relatifs à la réponse hypoxique. La définition des groupes repose sur la banque de données *MSIGDB* (version 2.5), publiée par les auteurs de la méthode *GSEA* [106]. Celle-ci comporte différentes catégories, présentées dans le tableau IV.B.6. L'analyse parallèle des trois jeux de données considérés, sur base de chacune des catégories représentées, pour chaque méthode, génère une quantité de résultats telle que leur description exhaustive sort du cadre de ce travail. A titre d'exemple, les paragraphes qui suivent s'attachent à comparer les résultats des méthodes sur deux sources de définition de groupes, pour le jeu de données E-MEXP-445.

Le critère « voie métabolique », a été choisi, afin d'illustrer la cohérence des résultats sur plusieurs sources de définition. Enfin, pour quantifier la cohérence des résultats des méthodes, nous proposons de caractériser la corrélation des résultats entre plusieurs jeux de données relatifs au même sujet expérimental.

Catégorie	Nombre	Description
C1.Positional	326	Localisation chromosomique
C2.BioCarta	249	Voies métaboliques (Biocarta)
C2.Canonical.Pathway	636	Voies métaboliques
C2.Chemical.Genetic.Perturbations	1168	Groupes définis en rapport avec des perturbations chimiques ou génétiques
C2.GenMAPP	138	Voies métaboliques (GenMAPP)
C2.KEGG	197	Voies métaboliques (KEGG)
C3.TF.targets	582	Groupes définis sur base des facteurs de transcription
C3.microRNA.targets	216	Cibles de microRNA
C5.GO.Biological.Process	791	Groupes impliqués dans les processus biologiques (GO)
C5.GO.Cellular.Component	212	Groupes impliqués dans la composition cellulaire (GO)
C5.GO.Molecular.Function	393	Groupes définis sur base de fonctions moléculaires (GO)
all	4323	L'ensemble des groupes étudiés

Table IV.B.6
Description des catégories de groupes définies au sein de la banque de données MSIGDB v 2.5. Le nombre de groupe représenté correspond uniquement aux groupes représentés au seins des 3 jeux de données envisagés, et dont le nombre de membres est supérieur à 2 et inférieur à 500.

IV.B.6.a. Analyse du jeux E-MEXP-445 : la réponse hypoxique au sein des monocytes

Le jeu de données E-MEXP-445 a été utilisé pour analyser plusieurs sources de définition de voies métaboliques, et de mettre en évidence la cohérence des résultats obtenus sur chaque catégorie de groupes, pour chacune des méthodes décrites. Nous avons choisi d'utiliser un seuil de sélection de 1% sur les *p-values*. A ce niveau de significativité, les différentes méthodes détectent un nombre variable de groupes de gènes. Certaines méthodes n'en détectent aucun. La table IV.B.7 dresse la liste de tous les groupes détectés, pour deux des catégories relatives aux voies métaboliques (C2.kegg et C2.biocarta), à titre illustratif. Les résultats associées aux catégories C2.genmapp et C2.canonical sont fournis dans l'annexe III.

L'examen de la table IV.B.7 montre que l'*ANOVA-2* détecte un très grand nombre de groupes, suivie par *GSA* basé sur la statistique *maxmean*, et sur la moyenne (*GSA.mean*). Ces méthodes, ainsi que nous l'avons décrit précédemment, sont particulièrement adaptées pour la détection de groupes de gènes dont les membres offre une réponse coordonnée, dans la même direction.

La méthode *FAERI* basée sur l'utilisation d'une distribution aléatoire détecte très peu de groupes. Une différence évidente apparaît pour la méthode *FAERI* évaluée sur base de permutations. Nous pouvons constater que certains groupes détectés par l'*ANOVA-2* sont également détectés par *FAERI* (permutations). Ceci illustre la capacité de la méthode à détecter également les groupes uni-directionnels. Les autres groupes détectés par *FAERI* (permutations) renforcent la description du contexte métabolique.

SAMGS et *GSEA* ne trouvent aucun groupe significatif à ce seuil. La méthode *GlobalTest*, quand à elle, ne détecte des groupes que si la significativité est évaluée sur base d'une distribution gamma.

Source	Méthode	Groupes
C2.biocarta	a2.fixed	actinypathway,chemicalpathway,crebpathway,erythpathway,feederpathway,glycolysispathway,malatexpathway,rhopathway,srcrptppathway,tall1pathway,vitcbpathway
	faeri.fixed.perms	raspathway
	GSA.mean.*	feederpathway,ghpathway,glycolysispathway,hifpathway,igf1pathway,plcpathway,sarspathway,vitcbpathway
	GSA.absmean.*	feederpathway,glycolysispathway
	GSA.maxmean.*	actinypathway,chemicalpathway,feederpathway,glycolysispathway,nkcellspathway,p53hypoxiapathway,plcpathway,sarspathway,vitcbpathway
	globaltest.gamma	blymphocytepathway,feederpathway,glycolysispathway,hifpathway,no1pathway,ptdinspathway
C2.kegg	a2.fixed	hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00061_fatty_acid_biosynthesis,hsa00071_fatty_acid_metabolism,hsa00072_synthesis_and_degradation_of_ketone_bodies,hsa00100_biosynthesis_of_steroids,hsa00190_oxidative_phosphorylation,hsa00232_caffeine_metabolism,hsa00330_arginine_and_proline_metabolism,hsa00380_tryptophan_metabolism,hsa00480_glutathione_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa00620_pyruvate_metabolism,hsa00630_glyoxylate_and_dicarboxylate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa00740_riboflavin_metabolism,hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa03320_ppar_signaling_pathway,hsa04150_mtor_signaling_pathway,hsa04370_vegf_signaling_pathway,hsa04530_tight_junction,hsa04612_antigen_processing_and_presentation,hsa05030_amyotrophic_lateral_sclerosis,hsa05040_huntingtons_disease,hsa05110_cholera_infection,hsa05120_epithelial_cell_signaling_in_helicobacter_pylori_infection,hsa05211_renal_cell_carcinoma
	faeri.fixed.null	hsa00140_c21_steroid_hormone_metabolism,hsa04940_type_i_diabetes_mellitus,hsa04950_maturity_onset_diabetes_of_the_young,hsa05210_colorectal_cancer,hsa05211_renal_cell_carcinoma,hsa05212_pancreatic_cancer,hsa05213_endometrial_cancer
	faeri.fixed.perms	hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00100_biosynthesis_of_steroids,hsa00511_n_glycan_degradation,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa04130_snare_interactions_in_vesicular_transport,hsa04210_apoptosis,hsa04620_toll_like_receptor_signaling_pathway,hsa04662_b_cell_receptor_signaling_pathway,hsa05040_huntingtons_disease,hsa05216_thyroid_cancer,hsa05221_acute_myeloid_leukemia
	GSA.mean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00052_galactose_metabolism,hsa00053_ascorbate_and_aldarate_metabolism,hsa00071_fatty_acid_metabolism,hsa00340_histidine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00710_carbon_fixation,hsa04660_t_cell_receptor_signaling_pathway,hsa05010_alzheimers_disease,hsa05050_dentatorubropallidoluysian_atrophy
	GSA.absmean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00031_inositol_metabolism,hsa00521_streptomycin_biosynthesis,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy
	GSA.maxmean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00100_biosynthesis_of_steroids,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00640_propanoate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa01510_neurodegenerative_diseases,hsa04660_t_cell_receptor_signaling_pathway,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05010_alzheimers_disease,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy,hsa05211_renal_cell_carcinoma
	globaltest.gamma	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00052_galactose_metabolism,hsa00272_cysteine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05211_renal_cell_carcinoma

Table IV.B.7 : Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Le seuil de sélection a été fixé à une valeur de 0.01 (p-values).

Les groupes de gènes détectés par les différentes méthodes peuvent être classés en quatre catégories :

- ☞ le métabolisme des sucres, le cycle de Krebbs et la phosphorylation oxydative ;
- ☞ les voies de signalisation : mtor, vegf, toll-like receptor, b-cell, t-cell, ppar... ;
- ☞ les pathologies : carcinome des cellules rénales, cancer colorectal, cancer de la thyroïde, maladie de huntington, diabètes, leucémie, alzheimer... ;
- ☞ les groupes liés spécifiquement à l'hypoxie : hifpathway, p53hypoxia... ;

La comparaison des groupes détectés avec le contexte biologique décrit, et entre les différentes méthodes, permet donc d'émettre les conclusions suivantes :

- ☞ les méthodes *ANOVA-2*, *FAERI*, *GlobalTest* (gamma), et *GSA* (*mean*, *absmean* et *maxmean*) détectent des groupes de gènes cohérents, et complémentaires, qui illustrent les perturbations du métabolisme en raison du manque d'oxygène ;
- ☞ les groupes spécifiquement liés à l'hypoxie sont détectés par certaines méthodes. Les autres méthodes les classent toutefois en bonne position (non représenté) ;
- ☞ plusieurs groupes qui sont également liés à un contexte hypoxiques sont détectés (cancers) ;
- ☞ plusieurs voies de signalisation détectées sont cohérentes avec les résultats décrits dans la littérature (VEGF, Toll-like receptor...).

Nous avons complété ces observations sur base de la liste des 20 groupes considérés les plus significatifs par chaque méthode, pour la catégorie C2.kegg (Annexe IV). Les observations réalisées en fixant un seuil sur les rangs fournit les mêmes informations que celles rapportées par la table IV.B.7. Elle se complètent toutefois par les observations suivantes : *SAMGS*, *FAERI*, et *GSA.maxmean* détectent plusieurs groupes liées à des voies de signalisation, à des pathologies, et au métabolisme de stéroides et lipides, alors que *l'ANOVA-2* est la seule à détecter la phosphorilation oxydative et présente un plus grand nombre de groupes liés au métabolisme des sucres, de même que *GSA.mean*.

Plusieurs observations générales peuvent être formulées sur base de tous les résultats obtenus :

- ☞ un très grand nombre de groupes détectés sont liés au métabolisme des sucres, des

acides aminés, des lipides, et à la phosphorylation oxydative, à la synthèse de l'ATP, en accord avec le contexte décrit, ainsi qu'à diverses voies de biosynthèses ;

- ☞ les résultats obtenus au départ des différentes définition des voies métaboliques sont cohérents, et décrivent soit les mêmes voies métaboliques, soit des voies métaboliques associées ;
- ☞ le nombre et la nature des voies métaboliques détectées diffère selon les méthodes.

Un examen approfondi de la nature des différentes perturbations sort du cadre de ce travail, qui vise simplement à explorer et améliorer la démarche analytique suivie, en tenant compte de la nature des différentes méthodes. Les diverses méthodes permettent de retrouver une partie différente de la vérité, de poser une regard différent sur les données, favorisant certains groupes selon les cas. Il apparaît évident, sur base de l'exemple biologique considéré, qu'il est préférable d'utiliser au moins deux types de méthodes pour avoir une vue d'ensemble des mécanismes impliqués. La méthode *FAERI* basée sur des permutations offre un bon compromis et est capable de détecter tous les types de groupes testés, mais l'interprétation des résultats est plus aisée par comparaison avec l'*ANOVA-2*, ou par la caractérisation des contributions individuelles des membres. Les méthodes *GlobalTest* (gamma) et *SAMGS* fournissent des résultats similaires à *FAERI*.

IV.B.6.b. Evaluation quantitative de la corrélation des résultats sur 3 jeux de données

L'hétérogénéité des groupes et la diversité des méthodes rendent complexe l'évaluation des performances des différentes méthodes sur des données biologiques, tel qu'observé au sein du paragraphe précédent. Il est toutefois possible de compléter notre évaluation par une comparaison des résultats, pour chaque méthode, sur plusieurs jeux de données relatifs au même thème. En effet, si une méthode est mise au point pour détecter des groupes impliqués, alors les mêmes groupes doivent être détectés sur des jeux de données relatifs à la même problématique.

La table IV.B.8 présente le nombre de groupes détectés pour la catégorie C2.kegg, sur les jeux de données E-MEXP-445, GSE-1056 et GSE-4086, ainsi que pour leurs intersections, pour des seuils de sélection placé à 0.01 et 0.05. Les méthodes *ANOVA-2*, *FAERI* (permutations) et *GlobalTest* détectent plusieurs groupes avec un niveau de significativité plus élevé, alors que *GSEA* et *SAMGS* n'en découvrent aucun. Pour le jeu de données GSE4086, un grand nombre de groupes sont détectés par les trois méthodes *ANOVA-2*,

FAERI et *GlobalTest*. La plupart des intersections observées avec le jeux de données GSE-4086 sont donc principalement dues à cet effet. Nous concentrerons donc l'analyse de la table IV.B.8 sur les jeux de données GSE-1056 et E-MEXP-445.

Pour un seuil de significativité de 1%, les méthodes qui détectent le plus de groupes communs aux deux expériences sont, dans l'ordre : *FAERI.permutations* (6 groupes communs pour minimum 14 groupes détectés), *ANOVA-2* (5/35), *GlobalTest.gamma* (1/8), et *GSA.mean* (1/14). Les autres méthodes ne fournissent aucune intersection. Lorsque le seuil de sélection fixé vaut 5%, l'ordre des méthodes est le suivant : *FAERI.permutations* (34/42), *GlobalTest.gamma* (20/45), *FAERI.null* (6/18), *ANOVA-2* (15/51), *GlobalTest.asymptotic* (4/26) et *GSA.mean* (1/8). Les autres méthodes ne fournissent aucune intersection. Ces résultats suggèrent que les groupes bi-directionnels détectés par *FAERI* et *GlobalTest* composent une part importante des groupes impliqués simultanément dans les deux jeux de données, par comparaison avec l'*ANOVA-2* et *GSA.mean*.

Pour quantifier équitablement la capacité des différentes méthodes à fournir le même résultat au départ de jeux de données différents, les coefficients de corrélation de Pearson ont été évalués sur les rangs associés aux groupes analysés, par comparaisons pairées des trois jeux de données, pour chaque méthode (Table IV.B.9).

Pour toutes les catégories envisagées, la méthode *FAERI.permutations* est associée aux coefficients de corrélation les plus élevés. A l'opposé, les méthodes *ANOVA-2*, *GSEA*, et *GSA.mean* fournissent des résultats peu corrélés. Les coefficients de corrélation obtenus pour les autres méthodes sont intermédiaires entre ces deux extrêmes, et varie selon la catégorie de groupes considérée.

0,010	a2.fixed	faeri.fixed.null	faeri.fixed.perms	GSA.mean.*	GSA.absmean.*	GSA.maxmean.*	globaltest.asymptotic	globaltest.gamma	globaltest.permutations	gsea.pval	gsea.fdr	samgs.pval	samgs.qval
emexp445	35	7	14	14	5	19	0	8	0	0	0	0	0
gse1056	36	9	51	5	0	8	1	15	0	0	0	0	0
gse4086	93	89	94	49	38	62	0	0	0	0	0	NA	NA
emexp445-gse1056	5	0	6	1	0	0	0	1	0	0	0	0	0
emexp445-gse4086	24	5	9	8	0	11	0	0	0	0	0	NA	NA
gse1056-gse4086	21	7	37	4	0	4	0	0	0	0	0	NA	NA
emexp445-gse1056-gse4086	4	0	3	1	0	0	0	0	0	0	0	NA	NA

0,050	a2.fixed	faeri.fixed.null	faeri.fixed.perms	GSA.mean.*	GSA.absmean.*	GSA.maxmean.*	globaltest.asymptotic	globaltest.gamma	globaltest.permutations	gsea.pval	gsea.fdr	samgs.pval	samgs.qval
emexp445	62	18	42	14	5	23	26	45	0	0	0	0	0
gse1056	51	27	80	8	2	12	33	70	40	11	0	15	69
gse4086	119	119	142	49	38	62	32	185	0	0	0	NA	NA
emexp445-gse1056	15	6	34	1	0	0	4	20	0	0	0	0	0
emexp445-gse4086	48	14	37	8	0	13	13	44	0	0	0	NA	NA
gse1056-gse4086	31	23	71	4	1	6	7	69	0	0	0	NA	NA
emexp445-gse1056-gse4086	13	4	31	1	0	0	2	20	0	0	0	NA	NA

Table IV.B.8 : Comparaison du nombre de groupes détectés dans la catégorie C2.kegg, et du nombre de détections communes, entre 3 jeux de données. Les résultats sont illustrés pour chaque méthode, pour des seuils de tolérance de 1% ou 5% d'erreurs. Les 3 jeux de données analysés sont E-MEXP-445, GSE-1056, GSE4086, et concernent la même thématique (la privation d'oxygène). La première colonne indique les jeux analysés. Les 3 premières lignes de chaque tableau caractérisent la *top-list* obtenue par chacune des méthodes d'analyse de groupes. Lorsque plusieurs jeux sont mentionnés en tête de ligne, le nombre de groupes rapporté dans les autres colonnes est l'intersection entre les résultats issus des différents jeux.

		emexp445	gse1056	gse4086	emexp445	C1.Positional gse1056	gse4086	emexp445	C2.BioCarta gse1056	gse4086	emexp445	C2.Canonical.Pathway gse1056	gse4086	emexp445	C2.Chemical.Genetic.Perturbati gse1056	gse4086	emexp445	C2.GenMAPP gse1056	gse4086	emexp445	C2.KEGG gse1056	gse4086	emexp445	C3.microRNA.targets gse1056	gse4086	emexp445	C3.TF.targets gse1056	gse4086	emexp445	C5.GO.Biological.Process gse1056	gse4086	emexp445	C5.GO.Cellular.Component gse1056	gse4086	emexp445	C5.GO.Molecular.Function gse1056	gse4086
a2.fixed	emexp445		3	12		0	7		0	19		0	24		6	14		0	26		0	26		0	4		1	7		10	8		0	14		3	9
	gse1056	3		4	0		0	0		0	0		1	6		4	0		14	0		2	0		0	1		0	10		4	0		8	3		9
	gse4086	12	4		7	0		19	0		24	1		14	4		26	14		26	2		4	0		7	0		8	4		14	8		9	9	
faeri.fixed.null	emexp445		36	0		15	0		17	29		28	19		18	13		21	11		46	25		30	0		45	10		30	3		53	22		53	20
	gse1056	36		25	15		6	17		32	28		33	18		24	21		38	46		30	30		0	45		33	30		35	53		57	53		30
	gse4086	0	25		0	6		29	32		19	33		13	24		11	38		25	30		0	0		10	33		3	35		22	57		20	30	
faeri.fixed.perms	emexp445		55	45		36	37		23	25		41	33		54	49		55	41		49	29		59	47		46	32		48	42		68	53		53	35
	gse1056	55		53	36		28	23		35	41		39	54		52	55		49	49		37	59		54	46		36	48		52	68		57	53		39
	gse4086	45	53		37	28		25	35		33	39		49	52		41	49		29	37		47	54		32	36		42	52		53	57		35	39	
GSA.mean.*	emexp445		3	11		0	7		0	13		0	19		8	8		1	42		0	23		0	7		6	7		0	9		1	3		6	17
	gse1056	3		4	0		9	0		0	0		0	8		7	1		0	0		0	0		12	6		10	0		1	1		0	6		0
	gse4086	11	4		7	9		13	0		19	0		8	7		42	0		23	0		7	12		7	10		9	1		3	0		17	0	
GSA.absmean.*	emexp445		17	22		4	18		5	2		7	19		12	14		1	39		12	23		0	11		23	21		13	21		41	43		32	27
	gse1056	17		24	4		3	5		25	7		24	12		15	1		25	12		21	0		23	23		15	13		27	41		37	32		33
	gse4086	22	24		18	3		2	25		19	24		14	15		39	25		23	21		11	23		21	15		21	27		43	37		27	33	
GSA.maxmean.*	emexp445		6	13		7	2		0	9		0	24		13	14		9	36		5	36		0	6		2	2		3	12		11	13		3	5
	gse1056	6		8	7		9	0		7	0		5	13		6	9		7	5		7	0		11	2		2	3		3	11		2	3		0
	gse4086	13	8		2	9		9	7		24	5		14	6		36	7		36	7		6	11		2	2		12	3		13	2		5	0	
globaltest.asymptotic	emexp445		16	22		0	20		6	23		14	35		15	17		14	52		14	25		1	17		14	23		16	19		26	4		13	19
	gse1056	16		14	0		7	6		19	14		18	15		10	14		15	14		15	1		22	14		8	16		12	26		29	13		9
	gse4086	22	14		20	7		23	19		35	18		17	10		52	15		25	15		17	22		23	8		19	12		4	29		19	9	
globaltest.gamma	emexp445		17	6		0	11		6	9		16	12		16	0		15	18		15	5		2	4		12	0		17	6		28	16		14	13
	gse1056	17		23	0		22	6		43	16		30	16		12	15		14	15		20	2		11	12		0	17		13	28		23	14		22
	gse4086	6	23		11	22		9	43		12	30		0	12		18	14		5	20		4	11		0	0		6	13		16	23		13	22	
globaltest.permutations	emexp445		15	10		0	21		1	2		11	21		17	6		17	48		9	14		1	0		7	10		17	2		31	2		9	17
	gse1056	15		12	0		4	1		18	11		18	17		10	17		13	9		22	1		20	7		1	17		9	31		13	9		7
	gse4086	10	12		21	4		2	18		21	18		6	10		48	13		14	22		0	20		10	1		2	9		2	13		17	7	
gsea.pval	emexp445		2	5		3	8		0	9		0	9		3	5		0	11		0	4		0	0		0	0		10	6		0	0		4	11
	gse1056	2		1	3		0	0		0	0		0	3		1	0		19	0		3	0		0	0		0	10		7	0		0	4		7
	gse4086	5	1		8	0		9	0		9	0		5	1		11	19		4	3		0	0		0	0		6	7		0	0		11	7	
gsea.fdr	emexp445		2	11		5	8		0	4		0	11		3	12		0	26		0	5		0	2		0	3		9	12		0	0		6	23
	gse1056	2		3	5		0	0		0	0		0	3		4	0		13	0		0	0		0	0		0	9		5	0		0	6		9	
	gse4086	11	3		8	0		4	0		11	0		12	4		26	13		5	0		2	0		3	0		12	5		0	0		23	9	

Table IV.B.9 : Coefficients de corrélation de Pearson, calculé sur les rangs, pour chaque méthode, entre 3 jeux de données, pour chaque source de définition des groupes. Les analyses ont été réalisées au départ des 3 jeux de données par l'ensemble des méthodes testées. Ensuite, pour chaque méthode, la liste des p-values est utilisée pour attribuer un score à chaque groupe (rang). Le coefficient de corrélation est calculé sur ces rangs, au départ des 3 jeux de données.

IV.B.7. Conclusions partielles

L'analyse de l'expression différentielle au niveau de groupes de gènes est une démarche rendue bien plus complexe que l'analyse individuelle, en raison de la diversité des critères biologiques considérés, et de la diversité des méthodes disponibles.

Les méthodes disponibles actuellement peuvent être classées en trois catégories : les méthodes de sur-représentation, les méthodes post-hoc qui reposent une stratégie d'analyse en deux étapes, et les méthodes dites globales. Les études présentées dans cette seconde partie du chapitre Résultats répondent aux limitations associées aux différentes méthodes.

Notre objectif repose sur l'utilisation d'une méthode multivariée de type *ANOVA-2*, pour améliorer les résultats de l'analyse de groupe. Nous proposons deux étapes supplémentaires, qui conduisent à la définition de la méthodologie *FAERI* (*Functional Analysis : Evaluation of Response Intensities*). *FAERI* a été développée pour permettre l'utilisation d'une seule méthode lors de l'analyse de groupes de gènes uni- et bi-directionnels. La statistique F^* associée a *FAERI* peut être évaluée soit sur base de données aléatoires, soit sur base de permutations d'échantillons (hypothèse auto-suffisante).

Après avoir présenté le modèle *ANOVA-2* envisagé et la méthodologie *FAERI*, une évaluation des performances a été réalisée, sur des données simulées, pour mettre en évidence la capacité de chacune des méthodes à détecter différents types de groupes (sur base de leur structure mathématique). Ces évaluations révèlent que la méthode *ANOVA-2* fournit les meilleurs résultats lorsque des groupes de gènes uni-directionnels sont analysés. Les résultats de la méthode *FAERI* sont toutefois proches des performances de l'*ANOVA-2*, et supérieurs aux autres méthodes. La méthode *FAERI*, de plus, atteint des performances supérieures à toutes les autres méthodes lorsque des groupes bi-directionnels sont analysés. Dans tous les cas, et pour toutes les méthodes, les performances des méthodes sont affectées lorsque les groupes analysés sont caractérisés par la présence de corrélation entre les gènes membres. *FAERI* s'avère néanmoins la méthode la plus appropriée pour ce type de groupes, ainsi que l'*ANOVA-2* dans le cas de groupes uni-directionnels.

Les évaluations réalisées sur base de données simulées ont été complétées par une analyse qualitative du jeux de données E-MEXP-445, relatif à la réponse hypoxique. Les résultats

obtenus montrent que l'*ANOVA-2* et *FAERI* (permutations) détectent à des seuils plus faibles les groupes les plus significatifs. Seule une partie des résultats obtenus ont été décrits, en raison de leur répétitivité. Ceux-ci affichent d'une part une grande cohérence vis-à-vis du sujet étudié (la privation d'oxygène), en révélant l'implication de groupes liés au métabolisme aérobie, à des voies de signalisation connues pour leur implication, à des pathologies au sein desquelles la privation d'oxygène joue un rôle. De plus, les analyses menées sur des groupes définis par des sources différentes présentent une grande cohérence et révèle les mêmes voies métaboliques.

Enfin, pour caractériser la capacité des différentes méthodes à classer correctement les groupes analysés, deux jeux de données supplémentaires, relatifs au même thème, ont été analysés. Le coefficient de corrélation de Pearson, calculé sur les rangs associés aux groupes, est proposé pour illustrer les performances des méthodes par comparaison entre les jeux de données. Ceux-ci montrent, quelle que soit la définition des groupes, que la méthode *FAERI* (permutations) est la plus apte à détecter les mêmes groupes au départ de jeux différents.

Au terme de ces évaluations, nous recommandons l'usage de deux types de méthodes lors de l'analyse de groupes de gènes, reposant respectivement sur une procédure uni-directionnelle (idéalement l'*ANOVA-2*, éventuellement *GSA.mean* ou *GSEA*) ou bi-directionnelle (idéalement *FAERI*, éventuellement *Globaltest.gamma* ou *SAMGS*). Toutefois, la méthode *FAERI* offre un bon compromis et est la plus appropriée pour l'analyse simultanée de tous les types de groupes envisagés.

IV.C. Modélisation et automatisation de l'analyse

Résumé

Ce chapitre présente l'automatisation des analyses que nous avons menées tout au long de ce travail.

Le postulat à l'origine du projet repose sur la modélisation des démarches suivies et présentées dans les deux premières parties du chapitre Résultats, et sur l'automatisation de l'analyse de l'expression différentielle, grâce à une utilisation optimale des méthodes récentes, en simplifiant l'usage et la paramétrisation des méthodes employées tout en permettant une configuration manuelle.

Nous proposons une stratégie d'analyse, en fonction des enseignements acquis en rapport avec l'analyse de gènes et l'analyse de groupes de gènes.

La stratégie proposée repose sur deux analyses parallèles, par gène et par groupes de gènes. Pour chacune de ces voies, plusieurs méthodes peuvent être utilisées. L'analyse individuelle peut être complétée par une analyse globale, grâce à la méthode *consensus* présentée dans la première partie des Résultats. La modélisation de l'analyse permet également d'envisager ces deux voies d'études dans un contexte de méta-analyse, utilisant alors le *consensus* pour résumer les résultats obtenus sur différentes sources. Enfin, le cas échéant, la stratégie suivie permet d'évaluer les performances des analyses réalisées, lorsque nous disposons d'une liste de gènes/groupes de gènes connus pour leur implication dans la problématique étudiée.

Nous montrons ensuite comment cette stratégie a été matérialisée au sein du package logiciel *PEGASE*, et de quelle manière son organisation interne le permet, avant de dresser la liste des méthodes disponibles pour chacune des étapes de l'analyse.

Le choix des indicateurs de performances proposés par *PEGASE* est présenté sur base de la stratégie d'évaluation des performances.

Les perspectives offertes par le logiciel *PEGASE* sont illustrées par deux exemples. Le premier exemple fourni est son interfaçage au sein du serveur en ligne *PHOENIX*, a mené à la publication conjointe de ces deux outils dans la revue *Central European Journal of Biology*, et dont le but est l'analyse de jeux de données publics avec des méthodes récentes pour en dégager davantage d'informations [19]. Le second exemple d'application présente un *benchmark* original, conçu comme extension à nos recherches pour faciliter l'évaluation des performances des méthodes sur base de données réelles. Nous illustrons la procédure proposée par une évaluation des performances maximales envisageables en utilisant la valeur attendue la variance pour évaluer la significativité.

IV.C.1. Introduction

Les objectifs poursuivis dans le cadre de ce travail sont multiples. Dès les premières études réalisées, l'orientation des recherches s'est voulue résolument exploratoire, combinant d'une part une évaluation comparative des méthodes d'analyse actuellement disponibles, et la volonté de croiser les données expérimentales avec les informations disponibles dans les bases de données publiques biologiques.

La première partie des résultats présentés porte sur l'analyse de l'expression différentielle des transcrits/gènes (selon le CDF utilisé) au départ de données issues de *microarrays*. Dans ce contexte, les différentes méthodes décrites dans la littérature ont été comparées, pour mettre en évidence une formulation statistique générale des méthodes basées sur l'amélioration de l'estimation de la variance, quel que soit le modèle statistique à l'origine des différentes méthodes (modèles Bayesiens empirique, estimateur de Stein, …). Pour un biologiste moléculaire, la diversité des méthodes existantes et de leurs interprétations post-analytique ne facilite pas le choix d'une méthodologie appropriée. Sur base des leçons tirées de la comparaison mathématique des différentes méthodes, nous avons montré que la complexité des développements statistiques a peu d'impact sur les performances des méthodes d'analyse. En effet, les meilleures méthodes reposent sur l'estimation d'une valeur moyenne de la variance, calculée sur base de l'estimateur individuel classique et d'un estimateur lié au « bruit de fond » (*background variance*). Le niveau de performance atteint repose essentiellement sur la méthode utilisée pour estimer ce second estimateur (médiane, relation moyenne-variance, coefficient de variation, …). Forts de ces observations, nous avons développé une nouvelle méthode, le *window t-test*. Reposant sur un modèle simple pour affiner l'estimation de la variance sur base de plusieurs *probesets*, la méthode développée permet d'atteindre le même niveau de performances que les meilleures méthodes.

Nous avons également observé que les résultats fournis par les différentes méthodes sont sensiblement différents, même si le niveau de performance atteint est similaire. Dès lors, nous avons proposé de combiner les résultats individuels sur base d'une approche *consensus*. Les performances atteintes par le *consensus* des méthodes sont comparables aux meilleures méthodes individuelles, dispensant l'utilisateur de choisir une méthode spécifique lorsqu'il ignore quelle méthode est la plus performante. De plus, les résultats obtenus stabilisent la relation entre la sensibilité et le taux d'erreur au sein de la liste des gènes sélectionnés, et améliore la comparaison entre des jeux de données différents.

Au cours de la seconde partie de ce projet, nos efforts se sont tournés vers l'analyse de groupes de gènes biologiquement reliés. La comparaison des différentes méthodes disponibles nous a permis de comprendre les avantages et inconvénients de chaque approche, et de cerner les étapes cruciales de l'analyse. Par extension, les critères déterminant l'efficacité des différentes méthodes ont été traduits en une stratégie dérivée de l'*ANOVA* à deux critères de classification croisées, la méthode *FAERI*. Nous avons montré que les performances atteintes par cette méthode simple surpassent les performances des méthodes actuellement disponibles, confirmant que la méthode *FAERI* constitue un choix judicieux pour une stratégie d'analyse de groupes de gènes.

IV.C.2. Modélisation de la stratégie d'analyse optimale

L'objectif fondateur du projet est de fournir une stratégie d'analyse optimale, automatisée, qui permette à un biologiste moléculaire, souvent dépassé par la complexité statistique des méthodes disponibles, d'obtenir les résultats les plus complets et les plus fiables possibles en toute simplicité. En terme conceptuels, l'approche suivie dans les deux premières parties de ce travail fournit une stratégie d'analyse représentée dans la figure IV.C.1.

Nous y avons représenté en rouge les étapes communes aux deux approches d'analyse de l'expression différentielle, centrées respectivement sur le *probeset* (analyse individuelle, représenté en bleu) ou sur un critère biologique reliant des gènes (analyse de groupes de gènes, représenté en vert). Dans les deux cas, la démarche modélisée s'initie par la préparation des données et la configuration de l'analyse.

L'analyse de l'expression individuelle repose sur plusieurs méthodes, dont les résultats sont introduits dans une procédure d'évaluation du *consensus* des résultats. Ainsi que nous l'avons démontré, une telle approche permet de se libérer du choix d'une méthode individuelle unique et fourni des résultats plus robustes et fiables.

La même stratégie d'analyse peut s'appliquer à l'étude de l'expression différentielle de groupes de gènes biologiquement reliés. Néanmoins, ainsi que nous l'avons montré, les meilleurs résultats sont fournis par la méthode *FAERI*, et l'évaluation du *consensus* de plusieurs méthodes d'analyse de groupes n'est donc pas justifiée (traits discontinus de couleur verte).

Les résultats de l'analyse de l'expression différentielle peuvent être comparés aux connaissances théoriques actuellement disponibles, en relation avec le sujet d'étude, afin d'en estimer les performances. Enfin, les résultats obtenus au cours des différentes étapes de l'analyse peuvent être illustrés à l'aide de tableaux et de graphiques divers.

Lors de l'évaluation des performances rapportée dans les deux premières parties de ce travail, nous avons également envisagé l'utilisation de plusieurs jeux de données différents relatifs au même sujet d'étude, pour comparer la robustesse des méthodes. Cette approche comparative constitue une démarche analytique biologiquement intéressante, appelée meta-analyse.

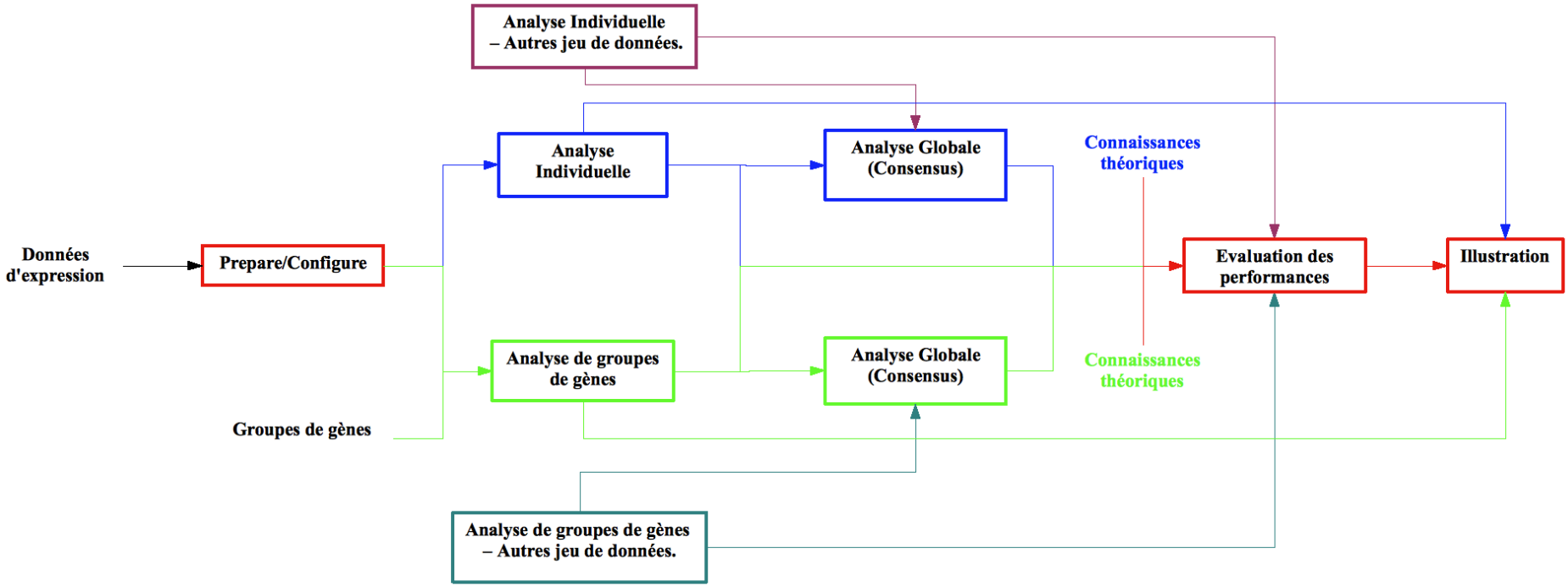


Figure IV.C.1 : Schéma conceptuel illustrant la stratégie d'analyse développée sur base des recherches présentées. L'analyse individuelle et l'analyse de groupes sont considérées comme deux tâches parallèles. Les résultats peuvent être évalués sur base de données externes, ou de l'analyse d'autres jeux. La méthode consensus permet d'utiliser plusieurs méthodes d'analyse pour obtenir des résultats plus fiables.

La figure IV.C.1 présente deux stratégies de meta-analyse complémentaires: d'une part, une méthode d'analyse statistique peut être utilisée en parallèle sur différents jeux de données, suivie par l'évaluation d'un *consensus* des résultats. D'autre part, les résultats obtenus sur un jeu de données précis peuvent être comparés aux résultats obtenus sur d'autres jeux de données, et fournir ainsi au chercheur une évaluation des performances de son analyse à la lumière des résultats obtenus ailleurs (validation). Les deux stratégies de meta-analyse sont applicables à l'analyse de l'expression individuelle et à l'analyse de l'expression de groupes de gènes, et sont représentées respectivement en bordeaux et en bleu clair dans la figure IV.C.1.

IV.C.3. Le package PEGASE

IV.C.3.a. Présentation du logiciel

Au fil des recherches menées dans le cadre de ce travail, nous avons écrit différents scripts permettant d'analyser les données issues de *microarrays* sur base de méthodologies existantes, réécrites au sein de notre unité afin de faciliter leur utilisation/paramétrisation. A partir de ces méthodes existantes, nous avons développé de nouvelles méthodologies (*window t-test*, *SAM* alternatif, *consensus*, *FAERI*), et effectué plusieurs tests de performance. Les procédures utilisées ont été rassemblées en plusieurs catégories de fonctions, en adéquation avec les différentes étapes de l'analyse présentées dans la figure IV.C.1. Le logiciel issus de cette intégration a été baptisé *PEGASE*, acronyme de « *Performance Evaluation And Global Analysis of Significant Expression* ».

Nous avons choisi de cibler en particulier deux types d'utilisateurs : les biologistes moléculaires, pour qui les statistiques sont bien souvent difficiles à appréhender, interpréter et illustrer, et les biostatisticiens/bioinformaticiens pour qui des processus d'analyse utilisant plusieurs méthodes en parallèle peut s'avérer utile, en particulier dans le contexte du développement de nouvelles méthodes et de la comparaison des performances des différentes méthodes.

Ces objectifs ont gouverné la stratégie adoptée pour la création du logiciel *PEGASE*, intégrant les différents aspects d'analyse de l'expression différentielle, l'automatisant, facilitant la paramétrisation, et fournissant des illustrations pertinentes des résultats obtenus.

Nous avons privilégié un aspect modulaire, décomposant chaque opération, chaque méthode, chaque analyse, en différentes étapes, avec pour but de pouvoir combiner les différentes étapes à volonté, ou pouvoir aisément substituer une étape à une autre pour créer de nouvelles méthodes. Cette approche s'est avérée particulièrement utile si nous considérons par exemple les différents méthodes dérivées du test de *t* de STUDENT, ou encore les différentes méthodes faisant intervenir des permutations, des étapes intermédiaires d'évaluation de paramètres, … A titre d'exemple, la méthode *window* a été conçue sur base d'une version embryonnaire du logiciel, en combinant plusieurs modules de calculs propres au test de STUDENT avec la correction de WELCH pour l'hétéroscédasticité, et au *regularized t-test* [11, 130, 143].

La figure IV.C.2 illustre l'organisation interne de *PEGASE*, par analogie avec le schéma conceptuel décrit dans le paragraphe précédent, synthèse des travaux effectués. Les fonctions présentées sont orchestrées par la fonction globale *pegase()*, qui assure la coordination des différentes étapes de l'analyse. Chacune de ces étapes est gérée par des fonctions spécifiques, utilisées successivement, qui organisent de manière optimale les procédures utilisées:

- ☞ Préparation (*pegase.na.rm*) et configuration de l'analyse (*pegase.conf*): les gènes pour lesquels un nombre insuffisant de données sont disponibles sont retirés du jeu de données (minimum deux valeurs dans chaque condition testée). La configuration repose sur le choix des étapes d'analyse, des méthodes envisagées à chacune de ces étapes, et des paramètres nécessaires à leur fonctionnement. La configuration peut-être effectuée manuellement ou automatiquement sur base de paramètres pré-définis, ou dérivés empiriquement au départ du jeu de données soumis à l'analyse.

- ☞ Analyse individuelle (*pegase.run*): sur base des méthodes sélectionnées, le jeu de données est soumis à plusieurs fonctions internes modulaires, combinées spécifiquement pour effectuer l'analyse individuelle en suivant l'une des procédures décrites par la communauté scientifique. La liste des méthodes disponibles est fournie par la fonction pegase.methods (STUDENT *t-test*, *window t-test*, *robust t-test*, *SAM test*, *regularized t-test*, *LPE test*, *Rank Products*, WILCOXON-MANN-WHITNEY *test*…)[11, 24, 77, 103, 130, 136, 143, 145].

- ☞ Analyse de groupes de gènes (*pegase.run.gs*): Lorsque l'utilisateur est en mesure de fournir une liste de groupes de gènes, ceux-ci peuvent être analysés avec les méthodes d'analyse de groupe disponibles (listées par *pegase.methods.gs*). A la différence de l'analyse individuelle, seules les méthodes *ANOVA-2* et *FAERI* font appel à des procédures internes. Les autres méthodes reposent sur des approches mathématiquement différentes, difficilement combinables, qui sont utilisées par *PEGASE* sur base de procédures externes développées par leurs auteurs respectifs (*GlobalTest*, *GSA*, *GSEA*, *SAM-GS*) [44, 53, 63, 106, 131].

- ☞ Analyse Globale (*pegase.cons*): Les résultats de l'analyse de l'expression différentielle sur base des différentes méthodes, qu'elle soit individuelle ou qu'elle corresponde à une analyse de groupe, sont utilisés pour évaluer le *consensus* des résultats. Le *consensus* peut être évalué au départ de résultats issus de différents jeux de données, dans un contexte de meta-analyse.

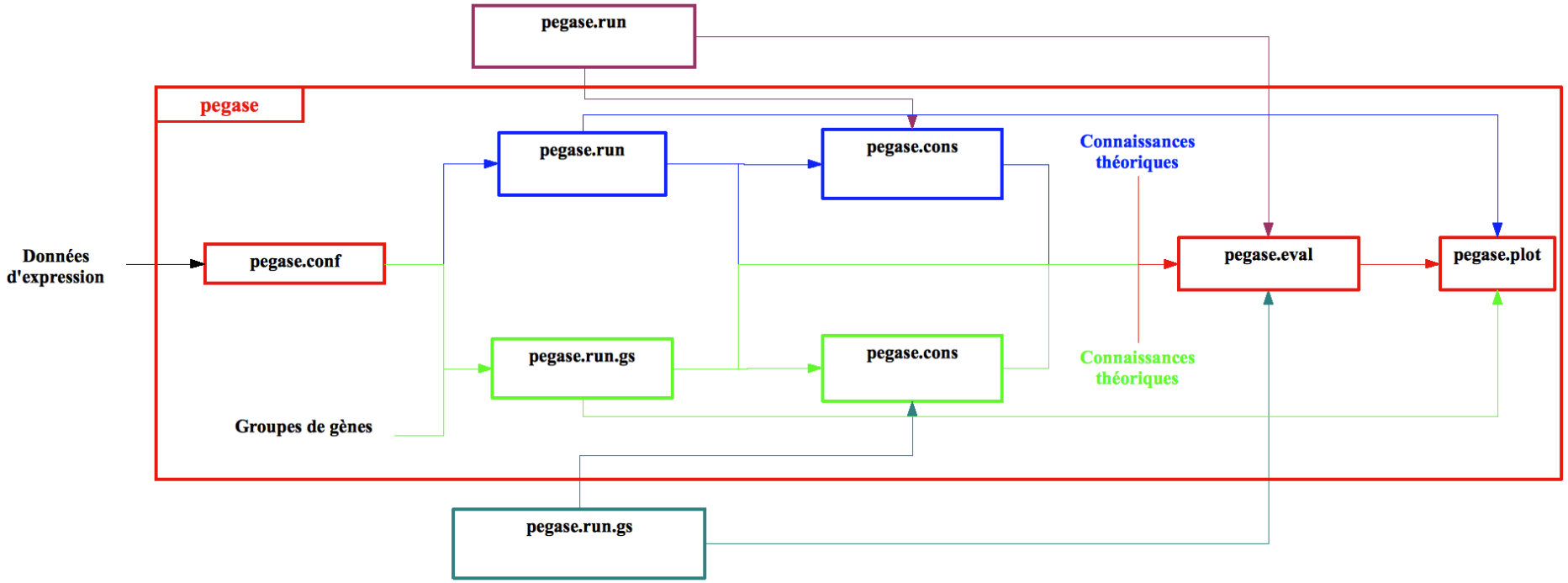


Figure IV.C.2 : Illustration de l'organisation interne de PEGASE, conçue pour correspondre au schéma analytique défini dans le cadre des recherches présentées.

- ☞ Evaluation des performances (*pegase.eval*): Lorsqu'une source externe permet de définir l'implication de gènes connus dans la problématique étudiée, cette liste de gènes est utilisée pour évaluer différentes statistiques représentatives de la qualité des résultats obtenus au terme des étapes précédentes. Les connaissances utilisables dans ce contexte sont soit liées au design de l'expérience (jeux de données *spike-in* ou simulés), soit aux connaissances théoriques issues de publications antérieures, soit aux connaissances empiriques obtenues en analysant un autre jeu de données. Les statistiques évaluées sont la sensibilité, la spécificité et le FDR, sur base de la table de contingence obtenue pour différents seuils de sélection.

- ☞ Illustration des résultats (*pegase.plot*): Les données, statistiques intermédiaires et résultats peuvent être illustrés de diverses manières pour vérifier la qualité de l'analyse et faciliter l'interprétation des résultats.

IV.C.3.b. Structure du logiciel PEGASE

Les deux paragraphes précédents ont permis de décrire le modèle conceptuel d'analyse de l'expression différentielle établi sur base de nos travaux, et les différentes étapes de l'analyse implémentées dans *PEGASE*. Le modèle conceptuel a été traduit en un ensemble de fonctions servant d'interface entre l'utilisateur et le logiciel, chaque étape conceptuelle ayant conduit à la création d'un fonction spécifique d'organisation des opérations effectuées. Ce modèle se doit d'être complété par un modèle complémentaire, illustrant schématiquement la structure du logiciel *PEGASE* (Figure IV.C.3). La première colonne de la figure correspond au modèle conceptuel d'analyse, décrit précédemment, qui forme l'interface utilisateur, ou *front-end*. Les autres colonnes présentent les fonctions qui effectuent les opérations nécessaires, le *back-end* du logiciel.

D'une part, la figure IV.C.3 étend le modèle conceptuel présenté, et décrit les étapes clés utilisées en analyse de l'expression différentielle. Toutes les méthodes d'analyse individuelles peuvent être décomposées de façon conceptuelle en trois étapes:

- ☞ l'estimation de paramètres au départ des données (variance, moyenne, rang, ...) ;

- ☞ le calcul d'un statistique individuelle ou d'un score au départ des estimateurs calculés ;

- ☞ l'estimation de la significativité du test au départ des statistiques individuelles.

Dans le cas de l'analyse de groupe de gènes, une étape supplémentaire apparaît avant le calcul de la significativité: le calcul d'une statistique de groupe au départ des statistiques individuelles.

D'autre part, plusieurs objectifs ont été poursuivis lors du développement de PEGASE. Ceux-ci peuvent être subdivisés en deux catégories, ayant trait respectivement aux objectifs scientifiques et aux objectifs informatiques. Du point de vue scientifique, PEGASE a été développé de manière à permettre une flexibilité de l'analyse, en choisissant la combinaison des étapes effectuées, et une flexibilité de paramétrisation qui autorise les bioinformaticiens à affiner leurs analyses sur base de leur propre expérience. Du point de vue informatique, nous avons également voulu permettre une automatisation de l'analyse et une auto-configuration du logiciel sur base de notre expérience et de caractéristiques empiriques du jeu de données choisi.

Enfin, la plupart des jeux de données disponibles concernent un grand nombre de données d'expression (22.283 pour la plateforme HG-U133A), si bien qu'une attention particulière a été portée sur l'optimisation des fonctions et de la structure de *PEGASE* afin de réduire au maximum le temps de calcul nécessaire.

Le jeu de couleurs utilisé dans la figure IV.C.3 illustre les différents types de fonctions implémentées dans *PEGASE* et la répartition conceptuelle des tâches, du point de vue informatique, pour rencontrer les objectifs visés:

- ☞ la gestion des données (bleu clair): les différentes étapes de l'analyse stockent les résultats dans une variable propre à *PEGASE*. Les fonctions de gestions des données permettent d'exporter les résultats dans un ensemble de répertoires et de fichiers, ou d'importer les résultats d'une analyse antérieure ;

- ☞ La gestion de la stratégie d'analyse (vert): les fonctions de préparation des données et de configuration de l'analyse se chargent de structurer la demande de l'utilisateur en une variable qui sera ensuite transmise aux différentes étapes de *PEGASE*.

- ☞ L'organisation d'une étape d'analyse (rouge): les fonctions de l'interface qui définissent l'analyse individuelle, l'analyse de groupe, et l'évaluation des performances peuvent être considérées comme des « chefs d'orchestres » dont le rôle est d'organiser les appels aux fonctions de calculs nécessaires pour utiliser chacune des méthodes statistiques implémentées.

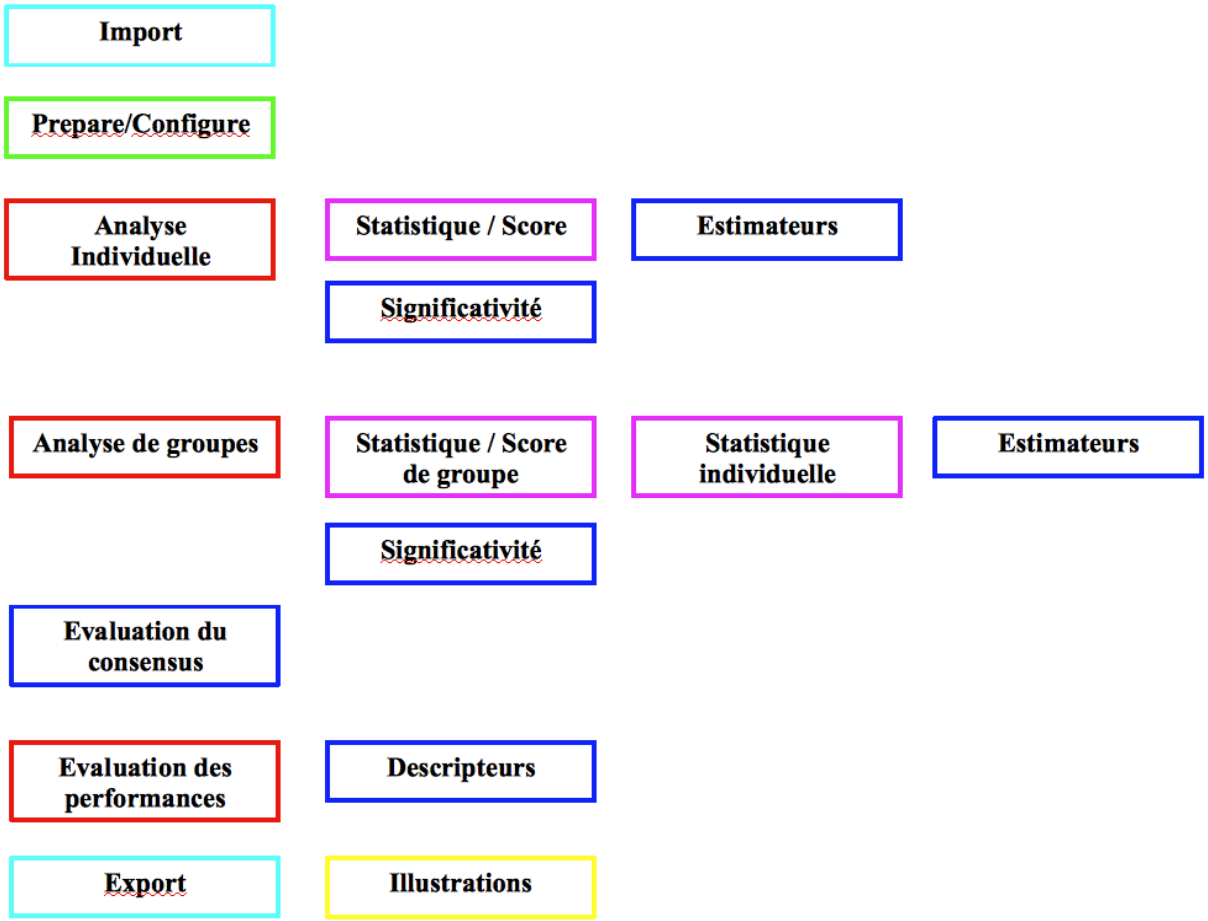


Figure IV.C.3 : Illustration, sous forme de table, de l'organisation interne de PEGASE, par correspondance avec le modèle d'analyse (1ère colonne), et avec les procédures statistiques décrites. En bleu foncé, les fonctions dites simples, qui ne nécessite aucune autre fonction. En magenta, les fonctions intermédiaires, qui font usage des fonctions simples et retournent leurs résultats vers des fonctions organisatrices (en rouge). Les fonctions d'import/export sont représentées en bleu clair.

- ☞ Les opérations simples (bleu foncé): les fonctions qui effectuent les calculs et fournissent le résultat à une fonction intermédiaire ou à une fonction organisatrice ;
- ☞ Les fonctions intermédiaires (magenta), effectuent des opérations simples, et exploitent d'autres fonctions de calculs. Ces fonctions sont typiquement associées au calcul des statistiques individuelles, ou de groupe, en organisant le calcul des estimateurs et en utilisant ces estimateurs pour calculer la statistique désirée ;
- ☞ Les fonctions d'illustration (jaune), utilisée pour traduire les résultats en graphiques et tableaux destinés à l'utilisateur, pour en faciliter l'interprétation.

La figure IV.C.3 ne présente qu'une représentation simplifiée de la structure de *PEGASE*. Un grand nombre de fonctions internes sont requises pour organiser et effectuer toutes les opérations. La description complète des appels de fonctions et de la hiérarchie structurelle de *PEGASE* ne sera pas présentée pour ne pas alourdir inutilement ce travail écrit.

IV.C.3.c. Description fonctionnelle de PEGASE

Le logiciel *PEGASE* a été conçu pour satisfaire au schéma analytique conceptuel présenté dans la figure IV.C.1. Nous avons décrit le rôle de chacune des étapes de l'analyse, ainsi que l'organisation générale du logiciel. A chacune des étapes de l'analyse, *PEGASE* offre différents outils d'analyse, présentés dans le tableau IV.C.1.

Analyse Individuelle **pegase.run**	*Analyse Globale* **pegase.cons**	*Analyse de groupes* **pegase.run.gs**	*Perf. evaluation* **pegase.eval**	*Illustrations* **pegase.plot**
Student t-test	Cons. p-value	ANOVA-2 fixed	Sensibilité	Sélection S0 (SAM)
Student t-test + Welch	Cons. p-value weighted	ANOVA-2 mixed	Specificité	d.stat/d.perm (SAM)
Window t-test	Cons. ranks	FAERI fixed (perms)	FDR = 1-Précision	p-values
Window Welch test	Cons. ranks weighted	FAERI mixed (perms)	p-value	4 figures perf.
Window Mixed test		FAERI fixed (null)	Nombre sélectionné	moyenne-variance
Regularized t-test		FAERI mixed (null)		fenêtre-réplicats
SAM test		GSEA (script)		
LPE test		GlobalTest (external)		
Robust Student test		GSA (external)		
Robust Welch test		SAM-GS (script)		
Wilcoxon ranksum test				
Rank Products				

Table IV.C.1 : liste des méthodes proposées par PEGASE pour chacune des étapes de l'analyse.

Les méthodes implémentées pour réaliser l'analyse globale ont été sélectionnées sur base des critères suivants:

- ☞ méthodes statistiques classiques: le test de STUDENT ainsi que sa correction pour l'hétéroscédasticité (WELCH) sont les outils statistiques traditionnels utilisés pour comparer deux séries de valeurs. Leurs équivalents robustes sont basés sur l'utilisation de la médiane et de la *MAD* en lieu et place de la moyenne et de l'écart-type lors du calcul de la statistique *t* [130, 143].

- ☞ méthodes statistiques classiques non paramétriques: l'équivalent non paramétriques du test de Student est le test de la somme des rangs décrit par WILCOXON et MANN & WHITNEY. Il repose sur l'utilisation des rangs des valeurs d'expression et la comparaison de la somme de ces rangs dans chacun des conditions comparées [103,145].

- ☞ méthodes de correction de la variance (modèles Bayesien empirique): plusieurs méthodes dérivent du test du STUDENT ou de WELCH, et proposent de corriger la variance sur base d'un modèle empirique Bayesien. Les représentants les plus largement utilisés de ces méthodes lors de l'initiation de ce projet sont les méthodes implémentées dans *SAM* (statistique *d*) et dans *CyberT* (*regularized t-test*). Le *regularized t-test* est l'une des méthodes qui atteint les meilleures performances d'après les tests que nous avons réalisés, en accord avec les travaux publiés dans la communauté scientifique [11, 136].

- ☞ méthodes basées sur l'exploitation de la relation entre le niveau d'expression et la variabilité: le *regularized t-test*, le test *LPE* et le test *window*, mis au point dans notre unité, sont 3 représentants de méthodes exploitant différemment la relation moyenne-variance. Le *window t-test*, méthode la plus simple qui utilise exclusivement l'estimateur « fenêtre », existe en trois déclinaisons, considérant une égalité ou inégalité des variances, ou un modèle mixte qui utilise un test d'égalité des variances et combine les deux approches [11, 18, 77].

- ☞ méthode non paramétrique destinée aux *microarrays*: la méthode du produit des rangs transforme le jeu de donnée en un ensemble de rapport de valeurs d'expression, le « *fold change* », pour chaque comparaison possible entre les deux conditions. La méthode affecte ensuite un rang à chaque gène au sein de la liste des *fold change* de chaque comparaison effectuée. Le produit de ce rang pour chaque

liste de *fold change* est utilisée comme estimateur représentatif de l'expression différentielle [24].

En regard de ces méthodes, la fonction *pegase.run* a été construite pour ne calculer qu'une seule fois les estimateurs utilisés en commun par plusieurs méthodes. La significativité est fournie dans tous les cas possibles pour répondre aux trois questions d'intérêt biologiques posées : quels sont les gènes différentiellement exprimés (bidirectionnels), sur-exprimés ou sous-exprimés (unidirectionnels) ?

La seule méthode actuellement implémentée pour l'analyse globale, étape que nous proposons d'ajouter à la stratégie analytique traditionnelle, est la méthode *consensus*. Ainsi que nous l'avons montré au cours de la première partie de ce travail, l'évaluation du *consensus* des résultats se décline en quatre versions qui fournissent des performances équivalentes, basées sur la *p-value* ou le rang de celle-ci, et avec pondération éventuelle des différentes méthodes utilisées lors de l'analyse individuelle.

La troisième colonne du tableau IV.C.1 liste les méthodes sélectionnées pour analyser l'expression différentielle de groupes de gènes biologiquement reliés. Parmi celles-ci, l'*ANOVA-2* a été sélectionnée comme représentant d'une méthode statistique classique. Ses conditions d'applications en font un bon candidat pour caractériser les différences d'expression unidirectionnelles de groupes de gènes (quels sont les groupes de gènes sur et sous-exprimés?). La méthode *FAERI* quant à elle, représente une méthode dérivée de l'*ANOVA-2*, optimisée pour réaliser des tests bidirectionnels (quels sont les groupes qui présentent le plus de différences d'expressions entre les deux conditions comparées?). Les quatre autres méthodes sont des représentants des méthodes actuellement disponibles pour l'analyse de groupe de gènes, basées sur 3 approches différentes : *GSEA*, méthode pionnière en deux étapes, basée sur l'utilisation d'un score d'enrichissement [106, 131] ; *GSA*, méthode en deux étapes basée sur une statistique de groupe (moyenne, valeur absolue de la moyenne, ou « maxmean ») [53] ; *GlobalTest*, méthode en une étape qui calcule une statistique de groupe au départ des données brutes [63] ; et *SAM-GS*, dérivé du test de T2 de Hotelling [44].

PEGASE offre également la possibilité de réaliser ces mêmes tests statistiques sur un jeu de données réduit en valeurs *Z* indépendamment pour chaque gène, et ainsi supprimer l'effet gène rapporté dans la seconde partie de ce travail. Il est important de noter que cette étape est intégrée au sein de la méthode *FAERI*.

Les méthodes *GSA* et *GlobalTest* ne sont pas intégrées à *PEGASE* et font l'objet d'appels à

des packages externes, fournis par leurs auteurs. La méthode *GSA* est proposée en utilisant les trois statistiques suggérées par le package, et la méthode *GlobalTest* peut être utilisée avec trois procédures d'évaluation de la significativité.

Les méthodes *GSEA* et *SAM-GS* sont également des méthodes externes, mais celles-ci nécessitent l'utilisation d'un fichier de script fourni par leurs auteurs (*GSEA*) ou d'un partie seulement du script fourni (*SAM-GS*), impliquant une préparation initiale du script par l'utilisateur avant d'être exploitable par *PEGASE*.

La description de l'évaluation des performances, ainsi que son illustration, sont traités dans le paragraphe IV.C.4. , et fourni une discussion sur les indicateurs utilisés.

IV.C.4. Evaluation des performances

IV.C.4.a. Introduction

L'analyse de l'expression différentielle est utilisée à plusieurs fins. L'évaluation des performances des méthodes se doit de rencontrer les objectifs initialement visés par l'analyse. A titre d'exemple, l'utilisation de *microarrays* dans un but diagnostique pour des pathologies spécifiques nécessite la connaissance de quelques gènes impliqués, avec très peu d'erreurs. A l'inverse, la description de voies métaboliques et des mécanismes mis en oeuvre au sein d'une pathologie nécessitent une détermination complète des gènes impliqués, ce qui implique un taux d'erreurs plus important.

La comparaison des performances des différentes méthodes statistiques développées dans le contexte des *microarrays* nécessite la détermination de plusieurs indicateurs, comme autant de point de vues portés sur les résultats. Ces indicateurs caractérisent la comparaison des résultats attendus avec les résultats observés. Plusieurs types de graphiques correspondent à des points de vues différents et doivent être interprétés en conséquence, pour illustrer correctement les performances des méthodes d'analyse de *microarrays*.

Lors de l'étude du profil d'expression, les *microarrays* sont typiquement utilisées pour mettre en évidence un certain nombre de gènes candidats, qui seront investigués plus avant avec des méthodes expérimentales plus fiables, comme la *qRT-PCR*. Généralement, en raison du coût, du temps nécessaire, de limitations techniques, le nombre de candidats sélectionnés pour la validation est généralement réduit. Dès lors, le seuil de sélection fixé par le biologiste moléculaire ne correspond pas à un seuil statistique défini sur la p-value, mais sur le nombre absolu de gènes sélectionnés.

PEGASE évalue les indicateurs associés à quatre types de courbes lors de l'étape d'évaluation des performances. Ces indicateurs peuvent être utilisés par les fonctions d'illustration des résultats pour générer les différents types de courbes décrits. Le logiciel comporte également une fonction de zoom, qui génère ces graphes en ne considérant que le début des courbes, plus informatif.

En complément à ces représentations graphiques, il peut être utile de résumer l'information visuelle en une donnée chiffrée, et le choix de l'indicateur utilisé doit être

représentatif des principaux critères de discrimination. Plusieurs fonctions ont été intégrées à *PEGASE* pour générer des tableaux récapitulatif des performances individuelles, sur base de ces derniers indicateurs.

L'évaluation des performances suppose que l'utilisateur possède une connaissance des résultats attendus, lié au design de l'expérience (jeux *spike-in*, jeux simulés, validations expérimentales, liste provenant d'une autre expérience). Elle peut être effectuée sur les résultats de l'analyse individuelle ou sur les résultats de l'analyse de groupes de gènes.

IV.C.4.b. La présentation graphique des performances

Les courbes ROC comparent la sensibilité et la spécificité des différentes méthodes (X: 1-spécificité ; Y: sensibilité). Ce type de figure est le plus répandu au sein de la communauté scientifique, bien qu'il ne permette pas de discriminer les méthodes comparées. En effet, le calcul de la spécificité fait intervenir un terme très élevé au dénominateur, le nombre de vrais négatifs, qui dilue l'information recherchée. En conséquence, toutes les méthodes comparées par ce biais fournissent des courbes superposées ou entre-croisées, proches de l'axe des ordonnées (Y) avec des performances apparentes équivalentes (GAIGNEAUX *ET AL.*) [60].

Dans le cadre de ce projet, nous avons proposé d'utiliser un autre type de représentation graphique des performances, dont l'interprétation est immédiate et intuitive. La sensibilité d'une méthode peut être formulée par « la proportion des gènes différentiellement exprimés qui sont détectés », ou, dans une formulation plus simple, « la proportion de la vérité qui est détectée ». La qualité de la détection, évaluée par l'erreur commise lors de la sélection des gènes sur base d'une méthode, peut être formulée par « le prix à payer pour trouver cette proportion », défini par le *FDR* (% erreur dans la sélection). Les coordonnées des figures présentées sont donc la sensibilité (ordonnée) et le *FDR* (abscisse).

Ce type de figure, que nous avons appelé *FDROC* (nous employons également le terme « courbes *ROC* modifiées »), permet une interprétation similaire à celle réalisée sur les courbes *ROC*: l'axe des ordonnées est conservé, et les méthodes les plus performantes fournissent des courbes plus proches du coin supérieur gauche (soit une détection maximale pour une erreur minimale).

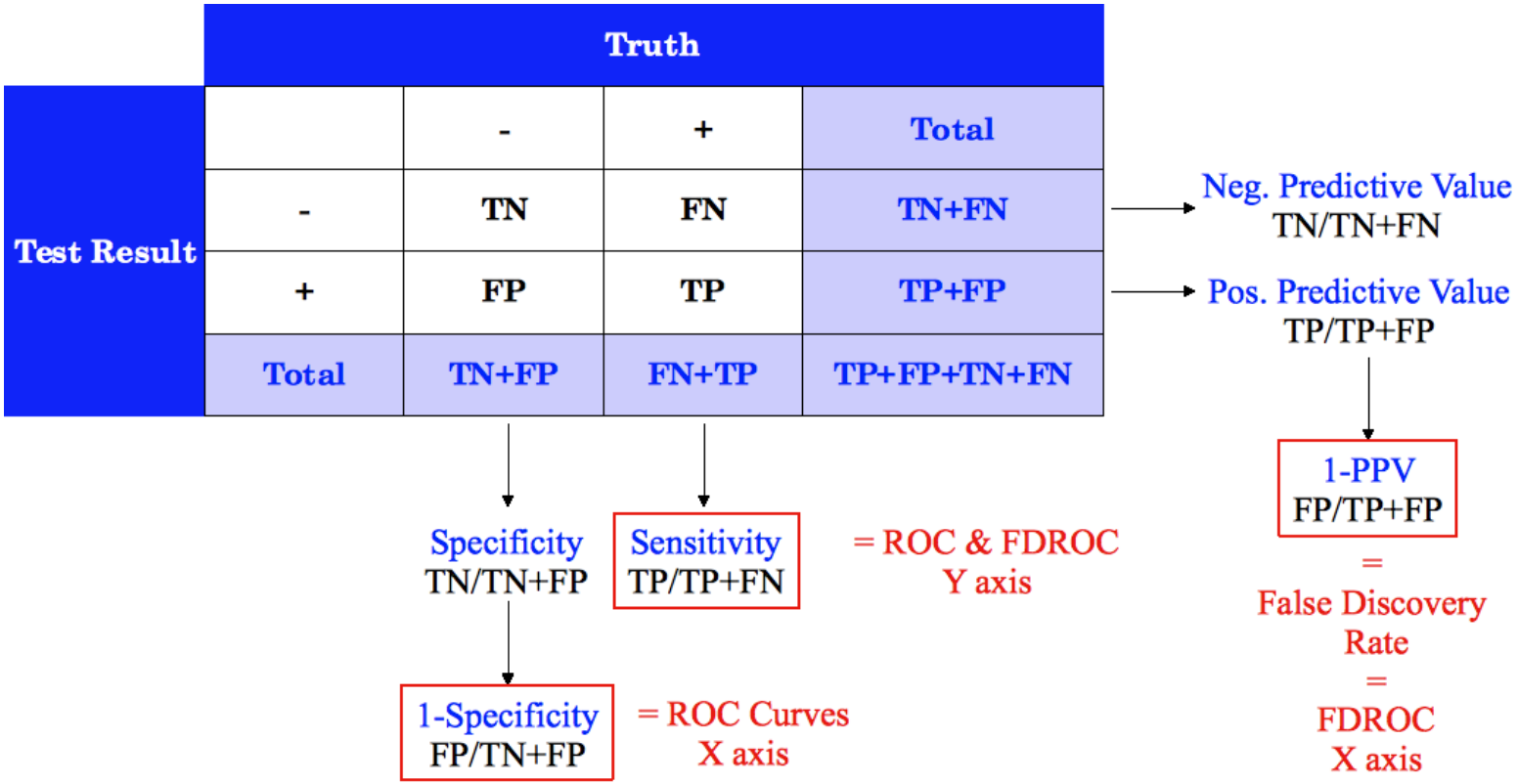


Figure IV.C.4 : Schéma qui illustre la correspondance entre les estimateurs utilisés pour les courbes ROC et FDROC, sur base de la table de contingence classique.

L'effet observé de « dilution » du nombre de faux positifs dans un océan de vrais négatifs, inhérent aux courbes *ROC*, n'a pas lieu d'être avec ce second type de courbes, qui est donc beaucoup plus discriminant. La figure IV.C.4 présente la correspondance de la table de contingence classique avec les indicateurs utilisés pour évaluer les coordonnées des courbes *ROC* et aux courbes *FDROC*.

Il est important également de mentionner qu'un autre type de courbe, dénommées « Precision/Recall Curves » *(PRC)*, compare la précision (ordonnée) à la sensibilité (abscisse). La formulation mathématique associée à la précision montre que ces courbes sont similaires aux courbes *FDROC* car la précision est le complément du *FDR* (*Precision = 1-FDR*), et évalue le taux de détections correctes au sein de la sélection. En théorie, les courbes *FDROC* et les courbes *PRC* présentent donc la même information, les axes des abscisses et ordonnées étant simplement intervertis, tout en remplaçant le *FDR* par son complément.

Une différence fondamentale distingue néanmoins les courbes *PRC* et les courbes *FDROC* : lors de l'évaluation des coordonnées de la courbe *PRC*, un lissage est effectué de sorte que la précision puisse uniquement diminuer lors du parcours de la liste des gènes. Les différents cas de figures rencontré au cours de ce projet nous ont montré qu'il arrive fréquemment que les premiers gènes détectés soient des faux positifs, auquel cas le début du parcours de la liste devrait présenter cet effet, qui s'accompagne d'une augmentation de la précision ou d'une diminution du *FDR* lorsqu'un plus grand nombre de gènes sont détectés. Les courbes *PRC* ignorent cet effet et imposent le plafonnement du niveau de précision associé au début du parcours de la liste. De plus, les biologistes s'intéressent typiquement à un nombre limité de gènes, et donc le départ de la courbe est un élément essentiel pour caractériser une méthode. A l'inverse, les courbes *FDROC* utilisées dans le cadre de ce projet, et adoptées comme référence au sein de notre unité, n'impliquent aucun lissage, car la stabilité du début du parcours est un élément fondamental de l'illustration des performances.

Enfin, deux autres types de représentations peuvent être utilisées pour caractériser les performances des méthodes. Toutes deux visent la caractérisation du taux d'erreur lors du parcours de la liste *(FDR)* en fonction du seuil de détection défini. La valeur seuil utilisée peut être de deux natures: la p-value, ou un nombre absolu.

Dans le premier cas, les performances des méthodes sont évaluées sur base de leur capacité à affecter une *p-value* réaliste au test réalisé. L'expérience acquise au cours de ce travail a

mis en évidence que les différentes méthodes fournissent une liste de *p-values* pour lesquelles la distribution varie selon la méthode. La définition d'un seuil sur base d'une *p-value*, et la comparaison des performances absolues nécessite donc une étape préalable de correction des *p-values*. Au cours de cette étape supplémentaire, chaque gène se voit attribué une valeur représentative du *FDR*, évaluée empiriquement sur base de la liste des *p-values*. Il s'agit d'un sujet d'étude à part entière, et aucune méthode actuelle n'estime correctement le *FDR*, à notre connaissance. La correction de la *p-value* reste donc un sujet délicat, mais n'intervient pas sur l'ordre dans lequel les gènes sont listés, et ne modifient donc pas les performances individuelles.

Le dernier type de représentation repose sur le nombre de gènes sélectionnés, et semble donc plus appropriée. De plus, le choix d'un nombre absolu de gènes correspond à une réalité expérimentale des biologistes moléculaires confrontés au coût et au temps de validation des résultats, qui conduit souvent à ne sélectionner que le nombre de gène qu'il est expérimentalement envisageable de valider [124].

IV.C.4.c. Quantification représentative des performances illustrées graphiquement

Afin d'illustrer quantitativement les performances des méthodes étudiées, il est utile de définir un indicateur représentatif des principales caractéristiques recherchées au travers de ces différentes illustrations. Dans le contexte de l'analyse des puces à ADN, sur base de nos recherches et en accord avec les préoccupations des biologistes moléculaires, ces critères sont les suivants:

- ☞ la sensibilité associée à la liste des gènes sélectionnés ;
- ☞ le risque d'erreur associé à la détection (qu'elle soit définie sur base du *FDR* ou de la spécificité) ;
- ☞ la stabilité associée au début de la courbe, qui concerne les premiers gènes détectés, et qui est représentative de l'évolution corrélée des deux premiers critères.

L'indicateur utilisé classiquement pour quantifier les performances des courbes *ROC* est l'aire sous la courbe (*AUC*), illustrée par la figure IV.C.5. Deux représentations fictives montrent que la même aire peut décrire plusieurs courbes *ROC*, qui pourtant ont une signification totalement différente en terme de performances ! D'une part, l'évaluation aboutit à la conclusion qu'il est possible de découvrir une partie de la vérité avec peu d'erreurs, et d'autre part, elle conduit à une découverte plus progressive de la vérité, avec

un taux d'erreur plus grand. Lors de la quantification des performances, l'utilisation de *l'AUC* masque cette différence d'interprétation et assigne la même aire aux deux courbes.

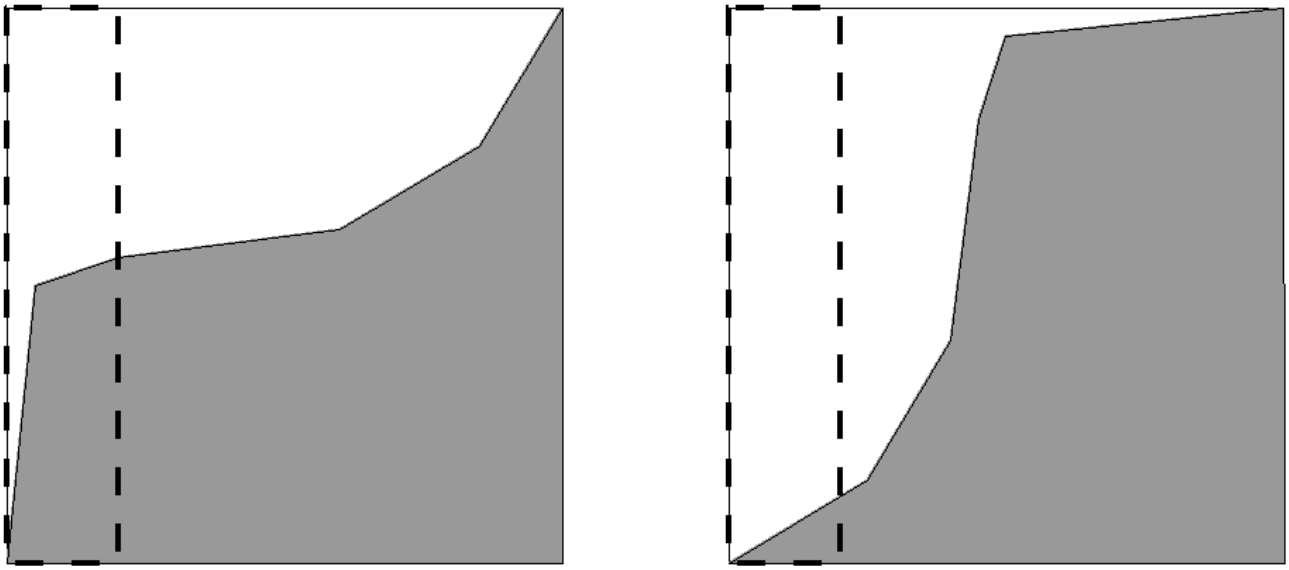


Figure IV.C.5 : Illustration de la limitation posée par l'utilisation de l'aire sous la courbe: les deux formes dessinées en gris représentent la même surface. La zone délimitée en pointillé concerne la région la plus importante du graphique. Cette zone, dans le cas des courbes *ROC* et *FDROC*, correspond aux gènes les plus significatifs. Sur la figure de gauche, la situation correspond à la découverte de 50% de la vérité avec peu d'erreurs. La seconde représentation, à droite, pourrait correspondre à une méthode moins performantes, qui ne découvrirait que 15% de la vérité pour un même taux d'erreurs.

En réponse à cette limitation des capacités de *l'AUC* à fournir un indicateur fiable vis à vis des 3 critères mentionnés, *PEGASE* fournit, en plus de cette indicateur, la possibilité de calculer celui-ci uniquement dans la région la plus sensible du graphique, qui correspond à un zoom réalisé sur la région du graphique associée au taux d'erreur le plus faible. A titre d'alternative, nous pensons que le choix de *l'AUC* peut-être remplacé par le niveau de sensibilité d'atteinte pour un niveau d'erreur choisi. Cette alternative trouve sa justification dans la procédure de sélection placée en aval des analyses : étant donné qu'un nombre limité de gènes est considéré, la qualité de la détection associée à des seuils de sélection plus strictes que celui utilisé n'influe en rien sur la qualité de la sélection effectuée pour des seuils moins strictes. De plus, considérant le problème posé par l'utilisation des courbes *ROC*, peu discriminantes en raison de la superposition courbes associées à chaque méthodes, il apparaît que *l'AUC* associé aux différentes méthodes est également peu discriminante.

Nous préconisons donc l'utilisation de *l'AUC* pour caractériser les courbes *FDROC*, plus à même d'illustrer la qualité de la *top-list*, et plus discriminante, et de n'évaluer que le début du parcours de la liste des gènes. Alternativement, l'utilisation de la sensibilité associée à

un niveau d'erreur choisi fourni une indication similaire des performances associées au début du parcours, en ignorant la stabilité de l'évolution de la sensibilité avec l'erreur commise.

Les résultats présentés dans la première partie de travail ont été illustré au travers de la représentation des performances sous forme de courbes *FDROC* (paragraphe IV.A.4.). A titre d'exemple, et en complément des résultats présentés précédemment, la table IV.C.2 présente les résultats obtenus sur base de l'*AUC* totale, ou partielle, comparativement au départ des courbes *ROC* et *FDROC*. Les mesures rapportées montrent que les différentes représentations chiffrées aboutissent aux mêmes conclusions, et que les méthodes de correction de la variance sur base d'une fenêtre fournissent les meilleurs résultats lors de l'analyse de jeux de données *spike-in*.

LS-95	***MAS 5***								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	98,69%	99,00%	98,98%	98,98%	98,98%	94,03%	**99,04%**	100%
	ROC20	19,21%	**19,47%**	19,44%	19,44%	19,43%	17,59%	**19,47%**	20%
	ROC10	9,39%	**9,63%**	9,61%	9,59%	9,58%	8,76%	9,62%	10%
	FDROC	50,76%	**79,40%**	70,73%	66,91%	71,12%	66,08%	72,76%	100%
	FDROC20	0,39%	**8,71%**	2,27%	0,35%	3,94%	8,20%	4,34%	20%
	FDROC10	0,17%	**1,79%**	0,55%	0,00%	0,18%	3,62%	0,17%	10%
	MAS 5 (LOG2)								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	98,71%	99,04%	**99,12%**	98,70%	98,37%	98,85%	98,35%	100%
	ROC20	19,21%	19,44%	**19,51%**	19,20%	18,93%	19,28%	18,90%	20%
	ROC10	9,38%	9,61%	**9,66%**	9,39%	9,23%	9,50%	9,21%	10%
	FDROC	63,13%	81,09%	**81,41%**	59,94%	65,46%	74,37%	65,24%	100%
	FDROC20	4,51%	**9,79%**	9,57%	0,95%	7,03%	5,48%	7,04%	20%
	FDROC10	0,83%	2,67%	**2,80%**	0,00%	1,64%	0,00%	1,65%	10%
	GCRMA								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	98,93%	98,97%	98,98%	**99,01%**	98,97%	98,72%	98,96%	100%
	ROC20	19,47%	19,57%	**19,58%**	19,51%	19,55%	19,40%	19,52%	20%
	ROC10	9,65%	9,73%	**9,74%**	9,69%	9,72%	9,65%	9,70%	10%
	FDROC	83,12%	85,43%	**85,64%**	82,69%	85,12%	80,76%	85,51%	100%
	FDROC20	9,83%	9,15%	9,14%	8,63%	9,50%	4,87%	**10,41%**	20%
	FDROC10	2,03%	1,25%	0,71%	0,96%	1,08%	0,00%	**2,07%**	10%
LS-133	***MAS 5***								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	97,15%	97,53%	**97,60%**	97,29%	97,24%	91,09%	97,23%	100%
	ROC20	18,58%	18,84%	**18,89%**	18,67%	18,66%	16,35%	18,67%	20%
	ROC10	9,10%	9,27%	**9,29%**	9,17%	9,16%	8,00%	9,17%	10%
	FDROC	73,81%	**84,74%**	82,18%	73,13%	78,60%	53,99%	79,40%	100%
	FDROC20	8,73%	**14,77%**	11,88%	6,98%	11,15%	7,79%	11,62%	20%
	FDROC10	2,62%	**6,58%**	3,87%	1,40%	3,87%	3,63%	4,17%	10%
	MAS 5 (LOG2)								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	97,04%	97,51%	97,53%	97,55%	97,49%	97,18%	**97,56%**	100%
	ROC20	18,47%	**18,82%**	18,81%	18,79%	18,74%	18,65%	18,76%	20%
	ROC10	9,01%	**9,25%**	9,23%	9,23%	9,19%	9,16%	9,20%	10%
	FDROC	71,53%	**84,01%**	79,38%	75,06%	77,43%	78,14%	77,20%	100%
	FDROC20	9,36%	**14,48%**	12,11%	8,09%	11,19%	9,72%	11,16%	20%
	FDROC10	3,53%	**6,52%**	5,22%	2,18%	4,48%	2,08%	4,52%	10%
	GCRMA								
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	95,39%	**96,30%**	96,18%	95,55%	95,50%	94,71%	95,50%	100%
	ROC20	17,88%	**18,49%**	18,37%	18,06%	18,05%	18,06%	18,05%	20%
	ROC10	8,76%	**9,14%**	9,06%	8,89%	8,88%	8,97%	8,88%	10%
	FDROC	79,55%	**85,94%**	84,32%	75,94%	82,92%	82,61%	82,66%	100%
	FDROC20	12,83%	**14,69%**	14,26%	10,54%	14,20%	12,05%	13,96%	20%
	FDROC10	5,12%	**6,00%**	5,78%	3,84%	5,91%	3,42%	5,69%	10%
Golden Spike 10c									
		Student	**Window t-test**	**Reg. t-test**	**SAM**	**Limma**	**LPE**	**Shrinkage t**	**Max. Value**
	ROC	87,18%	89,03%	88,96%	87,37%	88,51%	**94,91%**	90,12%	100%
	ROC20	14,10%	15,14%	15,12%	14,17%	14,49%	**17,18%**	14,99%	20%
	ROC10	6,33%	7,08%	7,04%	6,37%	6,54%	**8,10%**	6,84%	10%
	PRC	67,45%	75,75%	74,74%	67,98%	70,13%	**85,32%**	73,61%	100%
	PRC20	5,98%	10,62%	9,57%	6,22%	7,16%	**12,55%**	8,80%	20%
	PRC10	1,45%	4,45%	3,76%	1,60%	2,21%	**5,29%**	3,27%	10%

Table IV.C.2 : Comparaison des performances de plusieurs méthodes d'analyse individuelle sur les trois jeux de données *spike-in* LS-95, LS-133 et *Golden Spike*. L'aire sous la courbe *ROC* et sous la courbe *FDROC* ont été évaluées dans leur intégralité, ainsi que dans la région la plus informative du graphe (10% et 20%). Les deux meilleures méthodes sont représentées en gras ou sont soulignées, respectivement.

IV.C.5. Exemples d'utilisation de PEGASE

IV.C.5.a. Serveur PHOENIX: Intégration de PEGASE

Le projet de thèse de Benoît de Hertogh, mené parallèlement à ce projet, a permis de matérialiser l'approche préconisée au sein de notre unité en un serveur en ligne présentant une interface d'analyse de données de *microarrays*. Celle-ci est en relation avec une banque de donnée locale de jeux de données publiques, car nous sommes convaincus que les performances accrues d'une analyse optimale, combinant les meilleurs méthodes actuelles, nous permettra d'exploiter davantage les informations portées par les jeux de données publics, analysés souvent il y a plusieurs années, avec des méthodes moins performantes.

L'interface graphique a été baptisée *PHOENIX*, en accord avec la légende associée à cet animal mythique, et notre volonté de faire « revivre » les jeux de données publics. *PHOENIX*, à l'intersection de nos deux projets de thèse, permet d'importer un jeu de données ou d'utiliser un jeu présent dans notre banque de donnée locale, de le prétraiter (choix multiple et flexible), puis d'effectuer les étapes d'analyse de l'expression différentielle, analyse globale et évaluation des performances. D'autres fonctionalités seront ajoutées ultérieurement, relatives notamment au choix du fichier de définition, à l'analyse de groupes de gènes et à la méta-analyse.

Pour joindre nos deux thèse en un outil en ligne, la solution que nous avons adopté repose sur une répartitions des tâches de type *front-end / back-end*. L'interface traduisant les choix de l'utilisateur en un fichier d'instructions écrites en langage *R*, et qui reposent sur des appels aux packages *expresso*, *affy* et *gcrma* pour le prétraitement, et à *PEGASE* pour toutes les autres étapes de l'analyse.

Pour les différents formulaires de *PHOENIX* qui reposent sur l'utilisation de *PEGASE*, les options choisies sont simplement converties en paramètres transmis à *PEGASE*. Le script *R* résultant est ensuite soumis à une liste d'attente avant d'être exécuté sur le cluster de notre unité (80 coeurs, 64 bits).

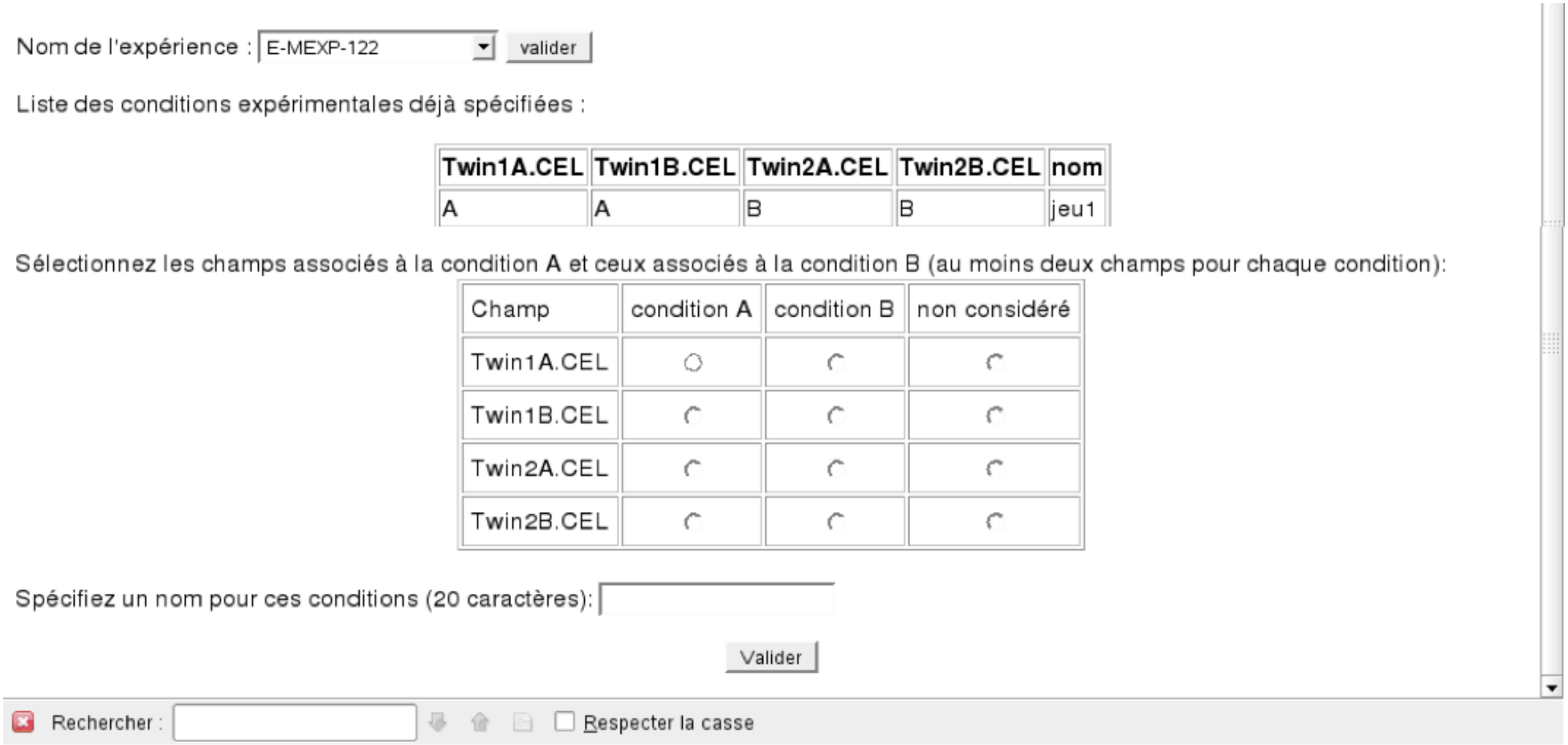


Figure IV.C.6 : Capture d'écran du formulaire de PHOENIX utilisé pour choisir ou définir le schéma expérimental étudié.

Les figures IV.C.6 et IV.C.7 présentent les onglets de l'interface *PHOENIX* qui concernent l'analyse de l'expression individuelle, entièrement réalisée par *PEGASE*. Dans une première étape, l'utilisateur est invité à choisir un jeu de données (importé ou issus d'une étape de prétraitement réalisée sur l'interface), et à en définir les conditions expérimentales. L'attribution d'un nom à cette définition expérimentale permet aux futurs utilisateurs d'utiliser le même schéma analytique, ou d'en définir un nouveau. Cette étape est fondamentale et ne peut être automatisée, d'une part car celle-ci n'est pas renseignée par les jeux de données, et d'autre part car il est souvent possible de définir plusieurs manière de regrouper les données.

La seconde étape, illustrée par la figure IV.C.7, permet ensuite d'analyser le jeux de données, avec la définition des conditions expérimentales adaptée, et permet à l'utilisateur de choisir une ou plusieurs méthodes d'analyse, éventuellement sur base d'une paramétrisation manuelle. Enfin, la dernière étape, optionnelle, permet à l'utilisateur d'évaluer les performances de l'analyse, en lui donnant la possibilité de fournir un fichier qui liste l'ensemble des *probesets*, et qui précise s'il est impliqué ou non dans l'expérience réalisée.

En complément à *PHOENIX*, nous proposons aux utilisateurs de télécharger *PEGASE* pour un usage local du logiciel. A cette fin, nous avons développé deux fonctions de gestion des données au sein de *PEGASE*. La fonction d'export des données permet, à chaque étape de l'analyse proposée par *PEGASE*, de stocker les données sous une forme directement utilisable par le scientifique (fichiers textes avec listes, et tableaux), mais également réutilisables par *PEGASE* via la fonction d'import des données, pour poursuivre l'analyse ou la reproduire en faisant varier certains choix méthodologiques. A l'avenir, *PEGASE* y inclura la création de graphiques et la genèse d'un rapport automatique, simultanément en fichiers *PDF* et dans le langage web *HTML*, directement utilisables par *PHOENIX* pour afficher les résultats. Au stade actuel, le script utilisé et les résultats sont stockés conjointement sur le serveur et peuvent être téléchargés par l'utilisateur, soit en tant que résultats, soit pour étendre ses recherches par usage local de *PEGASE*.

Afin de simplifier la navigation au sein des données générées par *PEGASE*, nous avons établit une correspondance entre la structure des répertoires créés et la structure de la variable utilisée par *PEGASE*. Chaque élément de la variable *PEGASE* correspond à un répertoire si l'élément possède plusieurs sous éléments, ou un fichier texte dans le cas contraire. Les fichiers textes comportent soient un tableau de valeurs, soit une série unique de valeurs. Les noms des répertoires/fichiers sont définis à l'instar de la variable *PEGASE*.

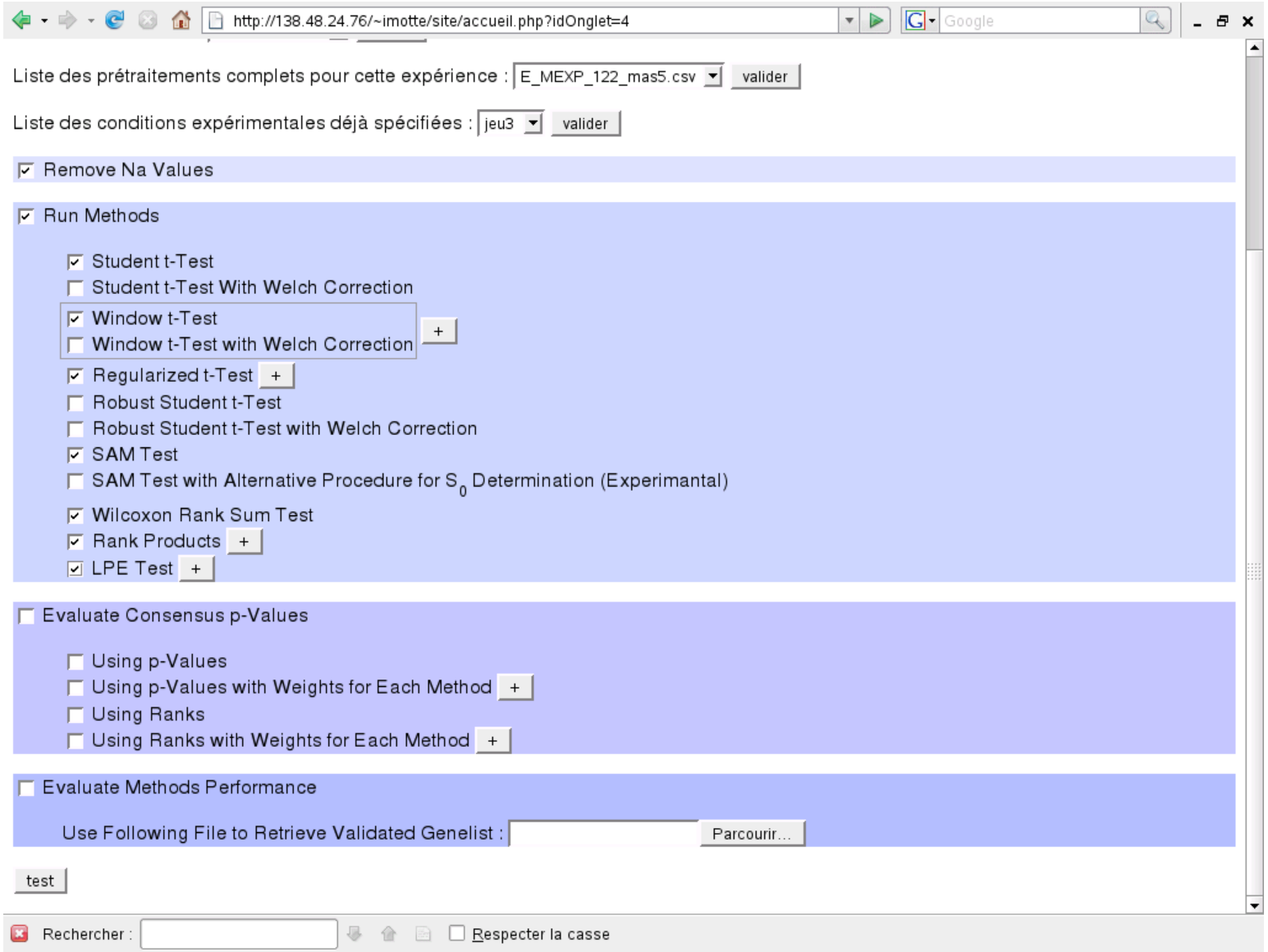


Figure IV.C.7 : Capture d'écran du formulaire intégré à PHOENIX pour définir la stratégie d'analyse souhaitée, et utilisée pour générer un script d'appel à PEGASE.

IV.C.5.b. Evaluation des performances sur des données réelles

L'évaluation des performances constitue une part importante des recherches présentées. Une évaluation adéquate des performances des différentes méthodes d'analyse de l'expression différentielle, au niveau individuel ou sur base de la définition de groupes de gènes, nécessite une connaissance préalable des résultats attendus sur un jeu de données connu. Un tel cas de figure n'existe pas à notre connaissance, si bien que la validation des méthodes développées, et la comparaison des performances, a dût être réalisée par plusieurs approches complémentaires. L'évaluation quantitative des performances a été réalisée sur base de l'utilisation de jeux de données réels, mais non biologiques, appelés « *spike-in* », pour lesquels des échantillons d'ARN en quantité connue ont été hybridés sur des biopuces. L'évaluation quantitative a également été réalisée sur base de simulations théoriques, ne reflétant donc pas la structure biologique rencontrée dans un jeu de données réelles. Plusieurs jeux de données biologiques réelles ont enfin été utilisés dans une analyse qualitative, pour illustrer la validité des résultats obtenus sur ces jeux.

Au cours de la mise au point des procédures d'évaluation, les recherches menées nous ont permis de mettre au point un nouveau jeux de données destiné à un usage systématique par les bioinformaticiens pour évaluer correctement les méthodes d'analyse. Ce projet, qui dépasse le cadre des recherches présentées ici, a été mis en oeuvre par Bertrand De Meulder dans le cadre de son travail de fin d'études et des premiers développement de sa thèse.

Le principe à l'origine de la création du jeu de données biologiques d'évaluation repose sur la construction d'une bibliothèque de jeux de données réels, caractérisés par un grand nombre de réplicats (>15), et sur l'utilisation d'une statistique indicatrice de la réponse observée sur ces jeux de données, pour chaque série de mesures. Partant du principe qu'avec un nombre suffisant de mesures, il est possible de caractériser chaque série de mesure avec un faible risque d'erreur, la vérité mesurée sur le jeu de données complet peut être comparée aux résultats obtenus sur des jeux de données de taille réduite (obtenu par échantillonage du jeu complet). La seconde originalité du *benchmark* réalisé repose sur la création de subsets dont le nombre de *probesets* est choisi par l'utilisateur, de même que la prévalence et le niveau de difficulté souhaité. Sur base de ces choix, le jeu de données est créé sur base des seuils statistiques qui traduisent ces critères, en construisant une librairie de *probesets* « différentiellement exprimés » et une librairie de *probesets* « identiquement exprimés ».

Les jeux de données simulés de cette façon répondent aux limitations actuellement posée par l'évaluation des performances, et permettent de comparer la vérité, construite de façon réaliste au départ de jeux réels, avec les résultats obtenus sur des jeux de difficulté croissante, reflétant la structure biologique des séries de mesures individuelle.

La validation de la procédure ainsi mise au point a été réalisée sur base de l'utilisation de *PEGASE*, dont l'automatisation a permis d'analyser parallèlement un grand nombre de jeux de données d'évaluation, sur base de la définition d'une difficulté croissante, et a ainsi permis l'étude de la robustesse des différentes méthodes d'analyse individuelle en regard du nombre de mesures.

Tirant parti de la naissance récente de cette procédure d'évaluation pour illustrer les applications possibles du package logiciel *PEGASE*, nous avons étudié le niveau de performances maximales accessibles. Toutes les méthodes récentes d'analyse de l'expression individuelle reposent sur l'amélioration de l'estimation de la variance, avec pour objectif d'obtenir un estimateur aussi proche que possible de sa valeur attendue. Cependant, il convient de se rappeler que chacune de ces méthodes utilise ensuite une statistique du type « t de STUDENT », laquelle est évaluée au départ des estimateurs de la moyenne (non corrigée), et de la variance (corrigée/stabilisée). Inévitablement, l'utilisation de plusieurs séries de mesures pour estimer plus justement la variance, n'a aucune influence sur les moyennes comparées, d'autant plus « variable » que le nombre de mesures est limité. Quelle est dès lors la limite entre ces deux effets antagonistes? Quelles performances pouvons-nous espérer atteindre sur base de la correction de la variance?

Pour répondre à ces questions, nous avons utilisé le jeu de données biologiques d'évaluation, et l'avons analysé avec une méthode « idéale » de correction de la variance. Chaque jeux de données généré à cette fin a été analysé en définissant une statistique *t* construite en utilisant les estimateurs classiques de la moyenne, et les estimateurs connus de la variance (mesurés sur le jeux complet). L'opération a été répétée pour plusieurs jeux de données générés avec un nombre différent de mesures, afin de caractériser la dégradation des performances due à l'erreur commise sur l'estimations des moyennes. La figure IV.C.8 présente les résultats obtenus, sur base d'un prétraitement des données réalisé avec *MAS 5.0* (log) ou avec *GCRMA*, pour une vérité « facile à trouver » et la figure IV.C.9 présente les résultats correspondants pour une vérité « difficile à trouver ».

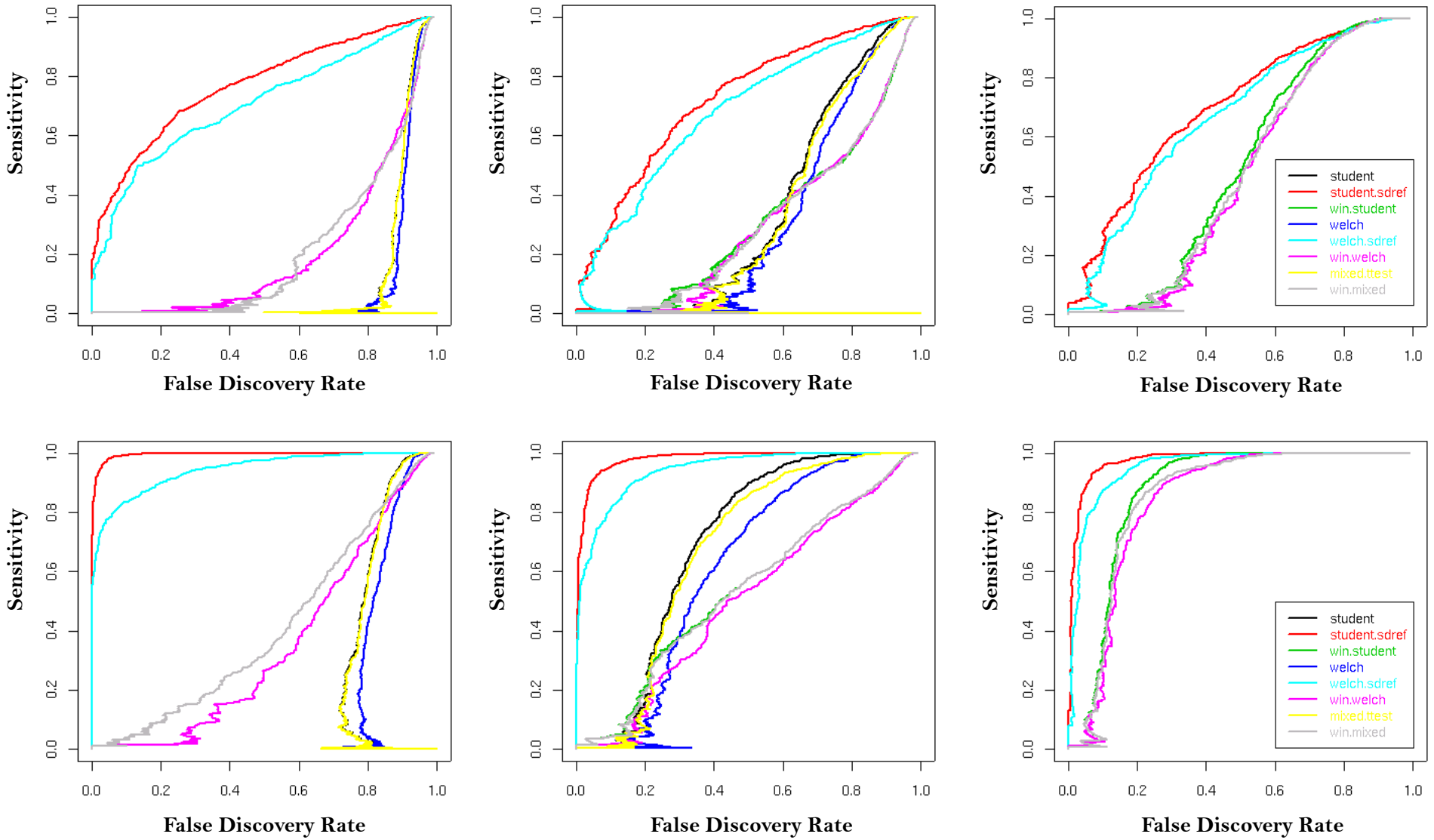


Figure IV.C.8 : Evaluation des performances d'une méthode de correction de la variance « idéale », sur base de la définition de la vérité dans un grand jeux de données, évalué ensuite sur des subsets. Les graphiques illustrés correspondent au prétraitement GCRMA (partie supérieure) et MAS 5 (log 2) (partie inférieure), pour des jeux de données de 2, 4, et 8 réplicats (de gauche à droite).

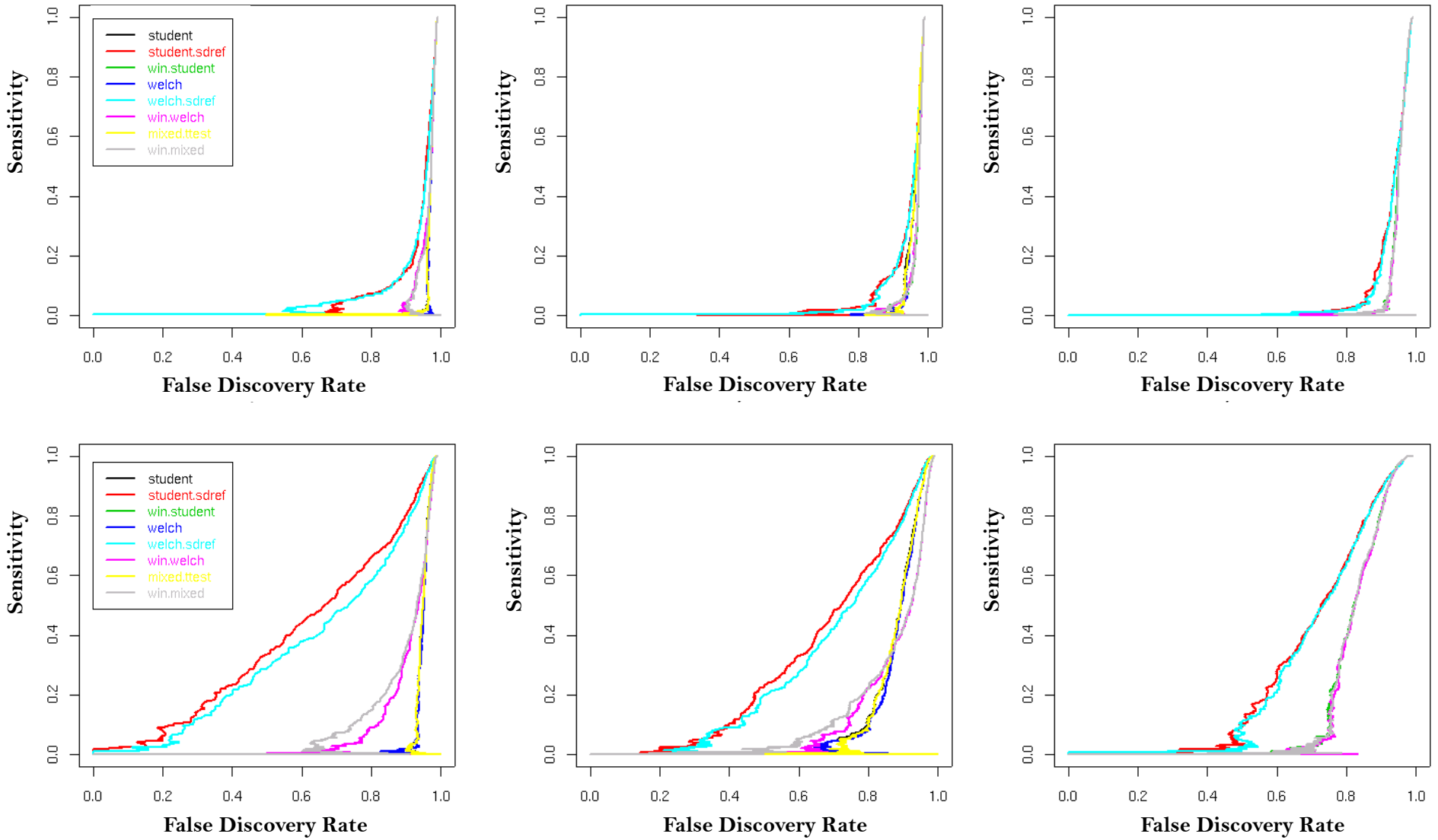


Figure IV.C.9 : Evaluation des performances d'une méthode de correction de la variance « idéale », sur base de la définition de la vérité dans un grand jeux de données, évalué ensuite sur des subsets. Les graphiques illustrés correspondent au prétraitement GCRMA (partie supérieure) et MAS 5 (log 2) (partie inférieure), pour des jeux de données de 2, 4, et 8 réplicats (de gauche à droite).

L'observation de ces figures révèle la différence de performance entre les méthodes d'analyse individuelles et le test de *t* réalisé avec la valeur réelle de la variance. Quel que soit le nombre de réplicats, le niveau de performance atteint par cette méthode « idéale » surpasse toutes les autres méthodes, et fournis des résultats comparables. L'erreur commise lors de l'estimation de la moyenne sur un jeu de données de taille réduite est donc négligeable en terme de performances. La différence de performances observée suggère que l'estimation de la variance est l'étape la plus sensible de l'analyse, et que les corrections apportées par le test *window* ne suffisent pas pour fournir un estimateur adapté. Il en va de même donc pour les autres méthodes basées sur une estimation corrigée de la variance (dont les performances sont comparables à la méthode *window*).

La méthode, dénommée *student.sdref* dans les figures IV.C.8 et IV.C.9 a été implémentée dans une version temporaire de *PEGASE*, préalablement à la réalisation de ces tests, afin de permettre à l'utilisateur de fournir une liste d'estimateurs de la variance à utiliser en combinaison avec le test de Student. La différence des performances observée offre la perspective d'analyser des jeux de données relatifs à un sujet d'étude commun, en utilisant les estimateurs de la variance du jeux de données le plus large pour analyser le jeu de données qui présente le moins de mesures. Nous suggérons donc de mener des analyses de ce type, sur des jeux de données biologiques bien décrits, afin d'en valider la démarche et de déterminer si cet échange d'informations améliore les performances en détectant davantage de gènes connus pour leur implication dans la problématique d'intérêt.

- La figure IV.C.8 illustre également un effet surprenant en ce qui concerne les performances de la méthode *student.sdref*. En effet, contrairement à toutes les autres méthodes, celle-ci voit ses performances diminuer lorsque le nombre de mesures augmente. Des tests complémentaires doivent être menés afin d'en identifier la raison. Il faut toutefois noter que l'évolution de ces graphiques avec le nombre de réplicats peut présenter une différence avec celle obtenue par échantillonnage d'un jeu de données classique. En effet, chaque jeu de donnée est simulé indépendamment, avec pour objectif de contrôler le niveau de difficulté. A l'inverse, dans un jeu de données biologiques, le nombre de réplicats influe sur le niveau de difficulté.

IV.C.6. Conclusions partielles

En débutant nos recherches, nous avons constaté que l'usage de méthodes peu performantes était très répandu (*fold change* et méthodes classiques utilisées avec un nombre insuffisant de résultats). Nous avons proposé durant la première partie de nos recherches, l'usage du partage d'informations entre les gènes, pour améliorer ces analyses, grâce au *window t-test*, publié dans la revue *Central European Journal of Biology*. D'autres méthodes récentes ont favorisé une démarche similaire à la nôtre, dont la paramétrisation repose sur d'autres critères [18].

Parmi toutes ces méthodes, le choix de la stratégie d'analyse optimale est une tâche difficile pour un utilisateur non statisticien. En réponse à ce constat, l'un des objectifs de nos travaux visait à fournir des outils d'analyse qui synthétise les enseignements tirés de nos recherches, pour guider et faciliter l'usage des méthodes envisagées pour les biologistes.

La première réponse que nous avons apportée à cet objectif a été présenté dans la première partie des résultats, et repose sur l'évaluation du *consensus* des résultats obtenus par différentes méthodes, associé à un niveau de performance comparable aux meilleures méthodes, sans identification préalable.

Nous avons décrit, dans cette troisième partie du chapitre Résultats, un schéma conceptuel d'analyse de l'expression différentielle, et qui résume l'ensemble des recherches réalisées sur ce thème depuis 2004. La stratégie proposée repose sur l'étude parallèle de l'analyse de l'expression des gènes et des groupes de gènes, sur base de plusieurs méthodes. L'utilisation d'un *consensus*, en analyse individuelle y figure comme étape additionnelle pour garantir la fiabilité des résultats obtenus. Dans un contexte de méta-analyse, la stratégie proposée permets d'envisager l'usage du *consensus* pour synthétiser les analyses réalisées sur des jeux de données propres à la même problématique. Enfin, lorsque les connaissances actuelles le permettent, ou lors de l'analyse d'un jeu de données simulé ou de type *spike-in*, le schéma d'analyse proposé intègre l'évaluation des performances, qui peut être également envisagée lors d'une démarche analytique de validation croisée des résultats.

Nous avons ensuite montré de quelle manière ce schéma analytique a été matérialisé et automatisé lors de la conception du logiciel *PEGASE*, et avons présenté une vue d'ensemble de sa structure interne, en accord avec le modèle défini, et conçu pour optimiser la combinaison des méthodes et indicateurs utilisés par les différentes méthodes. Grâce à cet outil, l'analyse des données peut être guidée automatiquement pour faciliter son usage

optimal par les biologistes, ou utilisée étape par étape avec une paramétrisation manuelle pour permettre aux bioinformaticiens d'adapter la procédure sur base de leur propre expérience, ou d'évaluer une nouvelle méthode sur base des fonctions d'évaluation comparative des performances.

Cette troisième partie des Résultats dresse ensuite la liste des étapes réalisée par *PEGASE*, et des fonctions implémentées ou utilisées au départ de scripts externes. *PEGASE*, à notre connaissance, est le seul outil actuel qui propose une stratégie d'analyse basée sur l'utilisation de plusieurs méthodes en parallèle. Cette démarche a ensuite été combinée avec notre objectif d'exploiter les méthodes les plus performantes pour extraire davantage d'informations à partir des données publiques, souvent réalisées avec des méthodes classiques. En intersection avec le projet de thèse de Benoît De Hertogh, nous avons ensuite présenté l'usage de *PEGASE* au sein du serveur en ligne *PHOENIX*, qui assure le lien avec les données publiques. *PHOENIX* permet de choisir un jeu de données et de le prétraiter, avant de définir automatiquement ou manuellement la configuration de l'analyse réalisée par *PEGASE*. *PHOENIX*, *PEGASE*, et la méthode *consensus* ont été acceptés récemment dans la revue *Central European Journal of Biology* [19].

Nous avons clôturé ensuite la présentation des résultats par un autre exemple d'extension de l'usage de *PEGASE*, pour faciliter l'évaluation comparative des performances sur base de données réelles. En interaction avec Benoît De Hertogh et Bertrand De Meulder, nous avons mis au point un outil de *benchmark* original, faisant usage de plusieurs jeux de données publiques pour créer sur-mesure un ou plusieurs jeux de données d'évaluation, en choisissant le niveau de difficulté souhaité. A titre d'exemple, nous avons présenté une évaluation des performances optimales qu'il est possible d'atteindre en corrigeant la variance, grâce à l'utilisation de la valeur attendue de la variance individuelle, pour chaque gène. Les résultats ont montré que les performances atteintes sont robustes en regard du nombre de réplicats, et largement supérieures aux meilleures méthodes actuelles, illustrant la possibilité d'améliorer davantage les performances de l'analyse individuelle si nous disposons d'un estimateur plus approprié de la variance.

Une extension intéressante à cette dernière évaluation consisterait à évaluer la variance individuelle, pour chaque gène, au départ d'un grand nombre de mesures collectées dans les données publiques. Pour faciliter des recherches ultérieures dans cette voie, nous avons inclus la méthode dans le logiciel *PEGASE*, pour permettre l'utilisation d'un jeu de données externe lors de l'évaluation de la variance individuelle.

V. Conclusions et perspectives

L'utilisation des biopuces permet de « photographier » le profil d'expression d'un génome complet au sein d'un échantillon et offre énormément d'applications, tant en recherches cliniques qu'en recherche fondamentales. Bien que la technologie soit largement décrite, et très répandue, elle est associée à une diversité de méthodes statistiques d'analyses qui peuvent désorienter le biologiste moléculaire, et fournir des résultats contradictoires.

La première partie de ce travail aborde la problématique de l'expression différentielle des gènes. La comparaison des différentes méthodes a été menée pour mettre en évidence le dénominateur commun des méthodes les plus performantes. Sur base de l'étude de la relation empirique entre le niveau d'expression et la variabilité individuelle, nous avons mis au point une procédure simple, le « *window t-test* », dont les performances peuvent être comparées favorablement aux meilleures méthodes, quelle que soit la procédure d'évaluation envisagée. Tenant compte des mesures associées à plusieurs gènes, a l'instar des meilleures méthodes, elle améliore l'estimation de la variance lorsque le nombre de mesures disponibles est limité. La correction apportée peut être formulée de diverses manières, rejoignant l'idée de la stabilisation de la variance véhiculée par les méthodes les plus efficaces, et peut également être envisagée comme une normalisation des données sur base du niveau d'expression atteint. La simplicité de la méthode, d'autre part, offre un avantage important en terme de temps de calcul, en raison du nombre de tests réalisés.

L'évaluation des performances des différentes méthodes révèle, en accord avec plusieurs publications récentes, que les méthodes les plus efficaces sont celles basées sur la stabilisation de la variance. Toutes partagent le point commun de corriger l'estimateur de la variance en partageant de l'information entre les les gènes. Le critère choisi pour déterminer le terme de stabilisation a un impact sur les résultats. Les quatre méthodes qui se distinguent des autres sont le test *window*, le *regularized t-test*, le *moderated t*, et le *shrinkage t* [11, 18, 109, 124]. Les deux premières reposent sur le partage d'informations entre les gènes qui partagent le même niveau d'expression. Les deux dernières partagent l'information sur l'ensemble du jeux de données. Deux autres méthodes partagent les informations entre les gènes: *SAM* et *LPE* [77, 136]. Les résultats obtenus avec ces deux méthodes sont plus variables, affichant de meilleures performances sur certains jeux, ou des résultats de moindre qualité sur d'autres. Les quatre meilleures méthodes, en revanche,

fournissent des performances similaires sur tous les jeux de données envisagés (simulés, *spike-in*, réels, simulés avec des données réelles). Dans tous les cas, ces méthodes ne sont efficaces que lorsque le nombre de réplicats est limité. Pour un nombre de réplicats plus importants, les performances du test de STUDENT augmentent, et les méthodes de stabilisations évaluent la même statistique que le test de *t*.

Il n'est pas toujours facile de choisir une méthode d'analyse, car l'absence d'indicateurs de performances sur des jeux de données inconnus limite nos capacités à déterminer quelle méthode est la plus appropriées à chaque jeu. A titre d'exemple, *SAM* est une amélioration du test de STUDENT qui repose sur l'estimation d'un terme correctif, mais la procédure employée conduit souvent à lui attribuer une valeur nulle, et par conséquent un résultat identique au test de STUDENT [136]. L'utilisation de permutations pour évaluer la significativité des résultats limite également la discrimination des gènes, particulièrement lorsqu'un nombre limité de mesures sont disponibles. L'utilisation de *SAM* ainsi que des méthodes qui utilisent des permutations pour attribuer une *p-value* doivent donc être considérées avec soin.

Partant du principe que les meilleures méthodes sont susceptibles de détecter les gènes impliqués, nous avons proposé l'ajout d'une étape supplémentaire au processus analytique: l'évaluation du *consensus* des résultats fournis par plusieurs méthodes. Les test réalisés montrent que les performances liées à ce *consensus* sont équivalentes aux meilleures méthodes, même si une méthode particulièrement inadaptée est utilisée pour l'évaluer. Les résultats obtenus sur les simulations nous incitent à utiliser cette étape additionnelle si nous ignorons quelle la méthode est la plus appropriée, garantissant la fiabilité des résultats fournis. Cette fiabilité a été évaluée sur le jeu de données E-MEXP-445, et montre la validité biologique des gènes détectés, en accord avec les études menées sur le même sujet (l'hypoxie). Plusieurs gènes candidats ont été proposés sur base de ces analyses.

Pour faciliter l'interprétation des résultats, la prise en compte de critères biologiques permet de donner du sens aux résultats obtenus, en considérant simultanément plusieurs gènes connus pour leur régulation commune, leur appartenance à une même voie métabolique, leur localisation chromosomique ...

Dans les premières études relatives à ce sujet, ces critères ont été considérés *a posteriori*, en aval de l'analyse de l'expression individuelle. Un seuil de sélection, fixé arbitrairement, fourni une liste de gènes candidats. L'annotation de ces candidats montrait alors que plusieurs d'entre eux sont liés, ce qui réduit la probabilité qu'ils aient été détectés par

hasard. Plusieurs études se sont alors donné pour objectif de quantifier cette probabilité, via des tables de contingences, par comparaison avec des groupes définis aléatoirement, ou sur base d'une théorie distributionnelle.

Ces études se limitent toutefois aux gènes les plus significatifs. Il n'y a pourtant pas de lien clair entre la force de la réponse mesurée statistiquement sur base des données d'expression, et leur impact biologique. Une protéine produite en plus grande quantité peut n'avoir aucun effet si son activité dépends d'une phosphorylation ou d'une autre modification post-traductionnelle. De même, des gènes faiblement exprimés peuvent avoir un impact considérable sur une voie métabolique, selon leur rôle. Plusieurs auteurs ont donc développé des méthodes d'analyse de groupes de gènes, utilisant la totalité des résultats de l'analyse individuelle, afin de détecter les groupes de gènes impliqués mais dont les membres présentent une réponse modérée (*GSEA* [106, 131], *GSA* [53] ...). Sur base d'une stratégie en deux étapes, plusieurs combinaisons de statistiques individuelles ont été envisagées.

Enfin, une troisième catégorie d'auteurs considèrent l'analyse de groupes de gènes comme une tâche parallèle à l'analyse individuelle, et ont développé des outils d'analyse qui utilisent directement les données d'expression, sans recourir à l'analyse individuelle (*GlobalTest* [63], *GlobalAncova* [103], *ANOVA-2*, *FAERI*).

Les recherches présentées dans la seconde partie de cet ouvrage montrent que les performances de ces méthodes sont liées à la définition des groupes. Plusieurs cas de figures ont été envisagés sur base de données générées aléatoirement, afin de sonder l'aptitude de chacune d'entre elles à détecter les diverses compositions possibles des groupes de gènes. Les résultats obtenus montrent que les critères les plus importants sont la taille du groupe, la direction de la réponse individuelle des membres, la corrélation, et la méthode utilisée pour attribuer la significativité. Les simulations réalisées en contrôlant ces critères montrent que les procédures mathématiques utilisées dans les différentes méthodes sont responsables du comportement observé, favorisant certains types de groupes, et pénalisant les autres.

Parallèlement à la comparaison des différentes méthodes, nous avons mis au point une approche analytique originale. Celle-ci repose sur une procédure similaire à l'*ANOVA-2*, mais implique trois modifications importantes de la stratégie. La première vise à réduire la variabilité due au niveau d'expression variable des membres, susceptible de donner plus de poids aux gènes les plus exprimés. Nous pensons que ce critère doit être ajouté aux autres critères mentionnés. La réduction des données en valeurs Z permet d'éliminer cet effet, et

de mettre chaque gène sur un pied d'égalité, tout en conservant l'information individuelle sur la force de la réponse (la statistique *t* individuelle est identique). La seconde correction apportée repose sur la définition des groupes de gènes. Biologiquement, une procédure de type *ANOVA-2* présente un intérêt non négligeable: elle permet de détecter des groupes pour lesquels les membres présentent une réponse dans la même direction. Ceci permet de détecter des groupes globalement sur-exprimés ou sous-exprimés. En revanche, si les membres sont tous impliqués, mais présentent des réponses opposées, une telle procédure considère que l'effet global est nul. La procédure *FAERI*, pour répondre à ce cas de figure important, utilise une réduction directionnelle pour cumuler les réponses individuelles. Enfin, la réduction directionnelle introduit un biais et la distribution nulle de la statistique évaluée est construite soit sur base de données aléatoires, soit sur base de permutations d'échantillons.

Les simulations réalisées montrent que, quelle que soit la structure du groupe étudié, les méthodes *FAERI*, *SAM-GS* et *GlobalTest* fournissent de meilleurs résultats que les autres méthodes. Une seule exception peut être mise en évidence: l'*ANOVA-2* détecte mieux les groupes unidirectionnels.

FAERI détecte les groupes testés à des seuils de significativité plus stricts que *SAM-GS*, mais leurs performances sont malgré tout corrélées. Ceci s'explique par la réponse similaire que *SAM-GS* apporte à l'analyse de groupes: elle est basée sur la somme du carré de la statistique individuelle. La direction de la réponse individuelle y est donc ignorée, seule la force de la réponse est prise en compte, sur base de la statistique *d*, similaire à la statistique *t* (et qui ne présente donc aucune dépendance vis-à-vis du niveau d'expression), et chaque gène est donc considéré sur un pied d'égalité avec les autres.

Les performances des autres méthodes, *GSEA*, et *GSA* avec les trois statistiques envisagées, fournissent des résultats différents. Leurs performances sont nettement plus faibles que les trois méthodes mentionnées précédemment, quelle que soit la composition du groupe.

Du point de vue de l'attribution de la significativité, et de la correction de celle-ci, les résultats montrent que les performances de *GlobalTest* varient selon la distribution envisagée, et que l'utilisation de la *q-value* par *SAM-GS* ou du *FDR* pour *GSEA* ont pour seul effet d'augmenter le nombre de faux positifs (*SAM-GS*) ou réduire le nombre de vrais positifs (*GSEA*).

Sur base des simulations, nous pouvons toutefois dire que toutes les méthodes détectent

correctement une partie du résultat, favorisant l'une ou l'autre composition de groupes, avec peu de faux positifs, pour les seuils envisagés.

Les méthodes d'analyse de groupes ont ensuite été critiquées sur base d'un exemple biologique concret: la réponse cellulaire à l'hypoxie. Dans un premier temps, l'analyse des groupes définis sur base des voies métaboliques révèle que l'ensemble des groupes détectés semblent en accord avec les implications biochimiques liées au manque d'oxygène, et à l'adaptation de l'activité cellulaire en conséquence. La méthode *ANOVA-2* et *GSA.mean* détectent davantage de groupes liés au métabolismes des sucres, suggérant que ces voies métaboliques présentent une réponse unidirectionnelle. D'autres voies, en revanche, apparaissent davantage représentées au sein des méthodes bidirectionnelles, et se réfèrent à des voies de signalisations diverses, dont certaines sont connue pour leur implication dans la réponse hypoxique. Enfin, plusieurs pathologies liées à l'hypoxie (dont plusieurs types de cancers) sont détectées. Les résultats obtenus sont donc cohérents. Les listes de groupes détectés au départ des différentes sources de définition des voies métaboliques concernent de plus les mêmes voies métaboliques, les mêmes pathologies, et les mêmes voies de signalisation.

Enfin, les corrélations de chacune des méthodes sur plusieurs jeux de données montrent que *FAERI*, basé sur les permutations, fournis des résultats plus reproductible que les autres méthodes. A l'opposé, *GSEA* et l'*ANOVA-2* fournisse des résultats non corrélés entre les diverse expériences. *GlobalTest* et *GSA.absmean* fournissent des résultats intermédiaires.

De tous ces résultats, ceux de la méthode *FAERI* sont particulièrement intéressants. Tous les tests réalisés montrent sa supériorité vis-à-vis des autres méthodes pour les questions envisagées. Il faut toutefois relativiser cette affirmation, car les test réalisés montrent des résultats similaires pour les méthodes *SAM-GS* et *GlobalTest*. L'utilisation de *FAERI* permet de détecter davantage de groupes avec un seuil de 0,01%. A ce seuil, *SAM-GS* et *GlobalTest* ne détectent rien sur le jeu de données biologique envisagé. Un seuil placé à 0,05 ne suffit pas non plus pour détecter des groupes avec ces méthodes, sauf *GlobalTest* s'il est évalué avec la distribution gamma.

L'analyse de l'expression différentielle des groupes de gènes semble donc fiable, à condition de tenir compte d'une part de l'hypothèse testée par les méthodes, des résultats obtenus avec les autres méthodes, et enfin, de la composition des groupes. A titre d'exemple, la délétion d'une région chromosomique sera avantageusement détectée par une procédure *ANOVA-2*, menée sur des groupes définis au départ de leur localisation chromosomique.

Des procédures telles que *FAERI*, *SAM-GS* et *GlobalTest* en sont également capables mais cette information y est diluée parmi des groupes de nature bidirectionnelle définis sur base d'autres critères. Il est donc important de tenir compte de la nature des groupes analysés, en regard des propriétés des méthodes, pour interpréter correctement les résultats d'une analyse de groupes.

Enfin, *GSEA*, la seule méthode hybride testée, présente un comportement différent et difficile à interpréter. Les groupes découverts sont associés à des performances inférieures à toutes les autres sur les simulations réalisées. Les résultats obtenus sur les trois jeux de données ne sont pas corrélés.

GSEA a cependant pour mérite d'exister et d'avoir été l'une des premières méthodes développées pour l'analyse de groupes de gènes, sur base de la liste complète des résultats de l'analyse individuelle, et a motivé plusieurs auteurs à étendre cette approche en tenant compte de critères biologiques ignorés par cette procédure.

L'étude de l'expression différentielle, individuelle ou sur base de critères biologiques connus, est donc un sujet complexe, nécessitant de nombreux choix méthodologiques, gouvernés par la structure des données, leur qualité, la nature du test réalisé, la manière dont la significativité est évaluée, et dans le cas de l'analyse de groupes, de la qualité de la définition des groupes et de la question à laquelle le biologiste souhaite répondre lorsqu'il entame ses recherches.

Avec pour objectif global d'améliorer la compréhension de l'analyse, et de fournir aux biologistes un outil capable de répondre au mieux à leurs questions, l'ensemble des enseignements portés par nos résultats ont été utilisés pour modéliser une stratégie optimale d'analyse, présentée dans la troisième partie. Celle-ci, en conjonction avec des motivations de flexibilité (pour les bioinformaticiens) et d'automatisation, a été matérialisée dans le package logiciel *PEGASE*. L'ensemble des travaux présentés ont été réalisés avec cet outil, au fur et à mesure des développement associés à chacune de ses versions embryonnaires.

PEGASE a été conçu pour effectuer séparément les analyses individuelles et les analyses de groupes de gènes, de même que l'évaluation des performances. Il est utilisé par le serveur web *PHOENIX* pour effectuer des analyses sur des jeux de données public. A ce titre, le duo *PHOENIX-PEGASE* (*front-end* et *back-end*) ont été mis au point pour tirer parti de notre expérience et permettre une analyse plus poussée de jeux de données publiés précédemment, avec un niveau de performance accru par la stratégie analytique que nous

avons décrit, basée sur l'utilisation de plusieurs méthodes d'analyse. A notre connaissance, PEGASE et PHOENIX sont les seuls outils actuellement décrits qui proposent d'évaluer un *consensus* sur base de plusieurs méthodes [19].

L'usage de *PEGASE* n'est nullement limité aux questions envisagées dans nos recherches, et l'usage des *window*, *consensus* et *FAERI* ne sont pas limités à l'analyse individuelle et à l'analyse de groupe.

Les outils développés dans le cadre de nos recherches présentent de nombreuses perspectives. Les extensions possible de ce projet sont nombreuses, et représentent chacune une amélioration de l'intégration des nombreuses connaissances biologiques et empiriques à notre portée.

La méta-analyse permet d'utiliser plusieurs sources de données et d'en dériver des informations globales. A ce titre, la démarche statistique employée est similaire à l'analyse de groupes de gènes. Ces deux thématiques ont pour point commun de considérer simultanément plusieurs séries de mesures, reliées entre elle. Ainsi, l'utilisation d'une méthode d'analyse de groupes sur un gène unique, au départ de plusieurs jeux de données, permet d'envisager plusieurs scénarios. La comparaison de jeux de données relatifs à la même problématique devrait être avantageusement réalisée par une procédure de type *ANOVA-2* pour détecter les gènes qui sont impliqués dans le sujet d'étude. En revanche, une procédure bidirectionnelle devrait favoriser la détection de gènes impliqués d'une manière différente dans deux pathologies, ou mettre en évidence l'action différente d'un médicament selon le sexe, le tissus, etc...

L'aspect méta-analyse peut également être envisagé par une extension de la stratégie d'évaluation d'un *consensus*. Au lieu de calculer le *consensus* sur base de plusieurs méthodes, celui-ci peut être évalué pour mettre en évidence les gènes impliqués dans plusieurs jeux de données. La même approche peut de plus être appliquée aux groupes de gènes, pour calculer le *consensus* des résultats obtenus sur différents jeux.

Ces différentes problématiques conduisent à envisager mathématiquement les notions d'intersections, d'unions, et d'exclusions. De telles questions sont fondamentales, et permettent de définir les mécanismes communs et les spécificités des divers jeux envisagés, des pathologies comparées... L'un des projets liés aux travaux présentés, confié à Michaël Pierre, repose sur l'utilisation de *PEGASE* et d'une stratégie méta-analytique originale d'évaluation de la significativité d'intersections et d'unions de résultats, appliqué à la thématique de l'hypoxie et des métastases, afin d'en dégager des enseignements sur les

mécanismes communs, et les valider par des techniques de biologie moléculaire.

Du point de vue méthodologique, la problématique de la méta-analyse est similaire à celle de l'analyse de groupes, et il serait très intéressant de mener davantage de recherches dans cette voie, afin de tirer avantage d'un nombre plus important de mesures. A notre connaissance, les approches suivies sont similaires à l'analyse de groupe ou à l'analyse individuelle, et les mêmes difficultés sont rencontrées. La plupart du temps, l'analyse est réalisée en deux étapes: une analyse individuelle est suivie par une analyse de « groupes de jeux ». Certains limitent l'analyse aux gènes les plus significatifs, au lieu de considérer la totalité des données. D'autres utilisent une statistique « moyenne » ou la somme des statistiques individuelles attribuée au même gène dans plusieurs jeux. Selon la statistique employée, ainsi que nous l'avons remarqué pour les méthodes d'analyse de groupes, certaines définitions de « groupes de jeux » devraient par conséquent être favorisées. Dans le contexte de la méta-analyse, l'impact du choix de la méthodologie favorisera donc préférentiellement les similarités, ou les divergences des réponses individuelles entre les jeux de données (informatives par exemple si on compare deux types cancers différents).

En terme méthodologiques, plusieurs autres développements peuvent êtres envisagés comme perspectives. La méthode *window*, par exemple, conceptualisée sous forme d'une normalisation, pourrait être utilisée en combinaison avec la méthode *FAERI*, pour développer une méthode *window-FAERI* qui commence par une étape de correction des données sur base du niveau d'expression. L'approche peut également être envisagée pour combiner n'importe quelle méthode de correction de la variance avec n'importe quelle méthode d'analyse de groupes. Les test réalisés par Marit Ackermann montrent cependant que l'utilisation d'une statistique « corrigée » comme statistique individuelle n'a pas une grande influence sur les résultats obtenus.

A l'inverse, la définition de groupes de gènes pourrait être utilisée pour améliorer l'estimation de la variance individuelle. Cette approche ouvre d'autres perspectives, telle que l'étude de la qualité de l'estimation de la variance en fonction du critère choisi pour définir les groupes de gènes.

Enfin, d'autres méthodes d'analyses de groupes doivent être développées, car il ne s'agit pas d'un sujet d'étude aussi linéaire que l'expression individuelle. Le type de méthode employé réponds à un type particulier de question, définie également par le type de groupe choisi.

Sur base d'un ensemble de critères biologiques, la méthode *FAERI* a été développée pour détecter les groupes qui affichent un comportement différent entre deux conditions.

D'autres solutions auraient pu être envisagées et constituent autant de perspectives à ce travail. A titre d'exemple, nous mentionnerons la possibilité d'utiliser un équivalent non paramétrique de la méthode *FAERI*, en remplaçant, pour chaque gène, les mesures d'expression par des scores. L'utilisation de scores en lieu et place des données d'expression devrait permettre une modélisation mathématique de la distribution nulle, et réduire ainsi le temps de calcul nécessaire pour évaluer la significativité.

Une démarche scientifique importante consiste à croiser les enseignements apportés par les différentes disciplines, en terme de connaissances, et en terme de stratégie. L'objectif des biostatisticiens est d'utiliser des démarches statistiques valides, adaptées à la question posée, et à l'information disponible. Il est dès lors surprenant de voir se répéter, à toutes les échelles d'analyse de l'expression différentielle, le développement de « nouvelles » méthodes, basées sur le même principe. A titre d'exemple, les informations des *probes* sont utilisées pour résumer l'information du *probeset* en une seule valeur. De la même manière, un « méta-gène » est évalué pour résumer l'information de plusieurs *probesets*, en méta-analyse ou en analyse de groupes, et le comparer sur base d'un simple *t* de STUDENT entre les conditions. A l'opposé, chacune des améliorations apportées consiste à évoluer d'une comparaison de valeurs uniques (le rapport des moyennes), vers la comparaison d'une série de valeurs (test de STUDENT, sur les valeurs d'expression, sur les meta-gènes, ...), pour finalement en déduire que les méthodes les plus efficaces utilisent l'ensemble des données, grâce à un modèle multivarié. Entre le moment ou deux valeurs résumées sont comparées, et le moment ou l'ensemble des données est considéré dans son intégralité, plusieurs années de recherches parallèles se sont écoulées dans les domaines de la méta-analyse, de l'analyse de groupes au départ des gènes, de l'analyse de gènes au départ des *probes*.

L'une des perspectives d'extension les plus porteuses pour améliorer significativement l'analyse biostatistique des données consisterait à intensifier les échanges inter-disciplinaires de solutions, en croisant les apports méthodologiques réalisés dans chaque discipline. De même que l'analyse de groupe peut être transposée à la méta-analyse, elle peut aussi l'être à l'analyse individuelle au départ des *probes*. Les questions biologiques posées à ces différents niveaux sont différentes, mais les réponses statistiques adoptées convergent. Les connaissances biologiques accumulées permettent une organisation hiérarchique des données (éventuellement sur base de critères différents) dont l'objectif est d'améliorer les analyses réalisées, quel que soit le niveau sur lequel se pose notre regard. Une telle organisation corresponds à des procédures statistiques multivariées, et il n'est pas surprenant de constater que les méthodes les plus performantes en analyse de groupes sont *FAERI*, *SAM-GS* (similaire au T2 de HOTTELING), et *GlobalTest*. Il n'est pas non plus

surprenant de constater que les méthodes d'analyse individuelle les plus efficaces reposent sur le partage d'informations entre les gènes.

Par extension, l'échange inter-disciplinaire pourrait être envisagé au delà de l'étude de l'expression, non seulement par corrélation avec d'autres sources d'informations (exon-array, séquençage, comparaisons inter-espèces), mais aussi et surtout par apport méthodologique. En croisant les méthodes d'analyses de groupes, d'analyse individuelle, et la phylogénie, nous pourrions comparer l'arborescence de GO, où chaque groupe de gène serait représenté par la *p-value* associée au test (et les gènes par les *p-values* de l'analyse individuelle). Différentes pathologies, comparées entre elles via ces arbres, ne révèleraient-elles pas plus efficacement leurs points communs et leurs spécificités ? Les outils disponibles actuellement étudient chacun des données différentes, à des échelles conceptuelles différentes. Un tel projet contribuerait à améliorer la compréhension intégrée des sujets étudiés.

Les connaissances actuelles de l'organisation génétique et fonctionnelle sont transcrites dans plusieurs bases de données (GO, KEGG, ...) très facilement utilisables en analyse de groupes. Il peut cependant être utile de définir les groupes sur base d'autres critères (organe, tissus, interactions protéiques, âge, sexe,...). A ce titre, je me réjouis de l'existence de la banque de données MSIGDB, dont le but est de lister des groupes de gènes en les classants selon différents critères. MSIGDB est une banque de données, qui permet l'apport humain de connaissances en dressant, entre autres, la liste des groupes mentionnés dans les publications, mais cette opération prends beaucoup de temps et les mises à jour sont peu fréquentes. Dans le cadre de ce projet de thèse, une autre voie de recherche a été initiée: la constitution d'une base de donnée, qui genère automatiquement des listes de groupes sur base d'un ou de plusieurs critères, collectés dans les principales sources d'informations biologiques. Une telle base de données permettrait, via une simple requête, de combiner plusieurs critères choisis pour générer une liste de gènes qui y répondent. Ce travail s'est révélé bien plus important qu'initialement imaginé, et a été repensé en profondeur. D'autres objectifs y ont été ajoutés comme la description des expériences, et la mise en relation des données empiriques avec les critères biologiques choisis. Le projet de base de données constitue à présent un projet à part entière, et a été confié à Eric Bareke.

Les connaissances actuelles peuvent être utilisées avec les outils récents d'analyse pour extraire de nouvelles informations au départ des données publiques. Une extension de ce projet consiste à appliquer systématiquement les outils d'analyse actuels sur les données disponibles, dans une approche méta-analytique. *PHOENIX* est né pour permettre

l'analyse des données publiques, en offrant, entre autres, une interface graphique à *PEGASE*. L'automatisation de l'analyse apportera différentes réponses, en connexion avec la base de données mentionnée précédemment, et facilitera la compréhension des mécanismes étudiés.

La visualisation des résultats, évoquée précédemment comme un arbre, peut être envisagée comme un réseau de voies métaboliques, dont les connexions sont définies par différents critères biologiques. Un projet d'étude de ces réseaux a été confié à Bertrand De Meulder.

VI. Matériel et méthodes

VI.A.
Jeux de données

VI.A.1. Introduction

L'analyse des performances d'une méthode suppose que nous disposons d'une connaissance préalable des résultats attendus de l'analyse de l'expression différentielle. Le jeu de donné idéal pour réaliser une comparaison appropriée des différentes méthodes d'analyse doit présenter un nombre significatif de mesures (nombre de réplicats), et être associé à une liste de gènes sur- et sous-exprimés.

A ce jour, aucun jeu de donnée de ce type n'a été décrit par la communauté scientifique. Plusieurs approches différentes ont donc du être suivies, sur base des jeux de données disponibles actuellement, pour évaluer correctement les performances sur base de différents critères. Traditionnellement, trois types de jeux de données sont utilisés pour comparer les performances :

- ☞ Les jeux de données « *spike-in* », dont le design expérimental correspond à une hybridation choisie d'ARN en quantité connue, sur différentes plateformes de *microarrays* ;
- ☞ Les jeux de données « simulés », qui se réfèrent à des données générées aléatoirement en lieu et place de mesures d'intensités, et qui tient compte de la structure d'un jeu de données réel ;
- ☞ Les jeux de données biologiques réels, parmi lesquels nous avons sélectionné d'une part un jeu de donnée avec un nombre élevé de mesures répétées, permettant une comparaison des résultats avec un nombre restreint de mesures, et les jeux de données relatifs à une thématique suffisamment décrites que pour nous donner un aperçu des résultats attendus, ou par validations croisées.

A ces trois approches classiques d'évaluation des performances, en collaboration avec Benoît de Hertogh, nous avons mis au point une quatrième approche d'analyse des performances, mise en oeuvre par Bertrand De Meulder et Benoît De Hertogh :

- ☞ Les jeux de données simulés au départ de données biologiques réelles.

L'approche suivie consiste à créer un jeu de donnée rassemblant plusieurs jeux de données biologiques comportant un grand nombre de mesures. Le jeu résultant est utilisé en conjonction avec un indicateur approprié pour générer des jeux de données de taille réduite (similaire à la statistique *t*), par la juxtaposition de mesures choisies parmi les gènes qui

affichent une différence d'expression et ceux qui participent au bruit de fond. Une telle stratégie permet de combiner les avantages des jeux de données simulés (connaissance des résultats attendus, contrôle de paramètres clés), et des jeux de données biologiques (variabilité individuelle non modélisée, mesures issues de données de réelles).

VI.A.2. Jeux de données « spike-in »

VI.A.2.a. Latin Square HG-U95 (LS-95)

Le carré latin mis au point par Affymetrix sur le modèle GeneChip HG-U95 comporte 14 expériences réalisé chacune avec 3 mesures. Pour chaque expérience, l'ARN relatif a 14 *probesets* ont été hybridés en quantité connues. La table VI.A.1 présente la définition des expériences réalisées, et la quantité d'ARN hybridée pour chaque *probeset* impliqué. Parmi les 14 expériences réalisées, 2 d'entre elles ont été répétées 4 fois (expériences M, N, O, P et Q, R, S, T, respectivement). Ces expériences ont été exclues de l'analyse pour assurer la cohérence des résultats. Nous avons également exclu l'expérience C, qui n'est représentée que par 2 mesures. Les 11 expériences restantes ont été utilisées pour évaluer les performances de plusieurs méthodologies sur base de 55 comparaisons disponibles.

Le jeu de données LS-95 est disponible publiquement sur le site web de la société Affymetrix [3].

VI.A.2.b. Latin Square HG-U133A (LS-133)

Le carré latin mis au point par Affymetrix sur le modèle *GeneChip* HG-U133A repose sur une stratégie expérimentale améliorée par rapport au modèle *GeneChip* HG-U95, et sur un modèle plus récent. La définition des 14 expériences envisagées repose ici sur 14 groupes de trois *probesets*, pour lesquels des quantités connues d'ARN sont hybridées sur les *microarrays*. Chaque expérience diffère donc des autres au niveau de 42 *probesets*. La table VI.A.2 présente la définition de chacune des expériences réalisées, et les quantité d'ARN hybridées. Chacune des expériences a été utilisée lors de procédures d'évaluation des performances, basée sur les 91 comparaisons rendues possibles par la stratégie suivie lors de la création du carré latin. Cependant, plusieurs *probesets* ont été ignorés lors de l'analyse, car ceux-ci étaient impliqués dans un phénomène d'hybridation croisée. Nous avons eu recours au package *AffyComp*, fourni par Affymetrix, pour retirer les *probesets* problématiques.

Le jeu de données LS-133 est disponible publiquement sur le site web de la société Affymetrix [4].

	37777_at	684_at	1597_at	38734_at	39058_at	36311_at	36889_at	1024_at	36202_at	36085_at	40322_at	407_at	1091_at	1708_at
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
A	0	0,25	0,5	1	2	4	8	16	32	64	128	0	512	1024
B	0,25	0,5	1	2	4	8	16	32	64	128	256	0,25	1024	0
C	0,5	1	2	4	8	16	32	64	128	256	512	0,5	0	0,25
D	1	2	4	8	16	32	64	128	256	512	1024	1	0,25	0,5
E	2	4	8	16	32	64	128	256	512	1024	0	2	0,5	1
F	4	8	16	32	64	128	256	512	1024	0	0,25	4	1	2
G	8	16	32	64	128	256	512	1024	0	0,25	0,5	8	2	4
H	16	32	64	128	256	512	1024	0	0,25	0,5	1	16	4	8
I	32	64	128	256	512	1024	0	0,25	0,5	1	2	32	8	16
J	64	128	256	512	1024	0	0,25	0,5	1	2	4	64	16	32
K	128	256	512	1024	0	0,25	0,5	1	2	4	8	128	32	64
L	256	512	1024	0	0,25	0,5	1	2	4	8	16	256	64	128
M, N, O, P	512	1024	0	0,25	0,5	1	2	4	8	16	32	512	128	256
Q, R, S, T	1024	0	0,25	0,5	1	2	4	8	16	32	64	1024	256	512

Table VI.A.1

Description du jeu de données Latin-Square HG-U95. Les valeurs fournies correspondent à la concentration des ARN ajoutés à l'échantillon hybridé. 14 expériences au total ont été mises au point avec 14 *probesets*. L'expérience C a été exclue de nos analyses car elle n'est représentées que par 2 mesures. Toutes les autres expériences ont été reproduites 3 fois. Les expériences M-P et Q-T ont également été exlues des analyses car elles sont identiques. Au total, 55 comparaisons pairées ont été considérées dans le cadre de nos recherches.

Group ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Gene ID	203508_at 204563_at 204513_s_at	204205_at 204959_at 207655_s_at	204836_at 205291_at 209795_at	207777_s_at 204912_at 205569_at	207160_at 205692_s_at 212827_at	209606_at 205267_at 204417_at	205398_s_at 209734_at 209354_at	206060_s_at 205790_at 200665_s_at	207641_at 207540_s_at 204430_s_at	203471_s_at 204951_at 207968_s_at	AFFX-r2-TagA_at AFFX-r2-TagB_at AFFX-r2-TagC_at	AFFX-r2-TagD_at AFFX-r2-TagE_at AFFX-r2-TagF_at	AFFX-r2-TagG_at AFFX-r2-TagH_at AFFX-DapX-3_at	AFFX-LysX-3_at AFFX-PheX-3_at AFFX-ThrX-3_at
EXP 1	0	0,13	0,25	0,5	1	2	4	8	16	32	64	128	256	512
EXP 2	0,13	0,25	0,5	1	2	4	8	16	32	64	128	256	512	0
EXP 3	0,25	0,5	1	2	4	8	16	32	64	128	256	512	0	0,13
EXP 4	0,5	1	2	4	8	16	32	64	128	256	512	0	0,13	0,25
EXP 5	1	2	4	8	16	32	64	128	256	512	0	0,13	0,25	0,5
EXP 6	2	4	8	16	32	64	128	256	512	0	0,13	0,25	0,5	1
EXP 7	4	8	16	32	64	128	256	512	0	0,13	0,25	0,5	1	2
EXP 8	8	16	32	64	128	256	512	0	0,13	0,25	0,5	1	2	4
EXP 9	16	32	64	128	256	512	0	0,13	0,25	0,5	1	2	4	8
EXP 10	32	64	128	256	512	0	0,13	0,25	0,5	1	2	4	8	16
EXP 11	64	128	256	512	0	0,13	0,25	0,5	1	2	4	8	16	32
EXP 12	128	256	512	0	0,13	0,25	0,5	1	2	4	8	16	32	64
EXP 13	256	512	0	0,13	0,25	0,5	1	2	4	8	16	32	64	128
EXP 14	512	0	0,13	0,25	0,5	1	2	4	8	16	32	64	128	256

Table VI.A.2

Description du jeu de données Latin-Square HG-U133-A. Les valeurs fournies correspondent à la concentration des ARN ajoutés à l'échantillon hybridé. 14 expériences au total ont été mises au point avec 14 groupes de 3 *probesets*. Au total, 91 comparaisons pairées ont été considérées dans le cadre de nos recherches.

VI.A.2.c. Jeu de données « Golden Spike »

Le jeu de données de Choe est assez inhabituel, et présente un intérêt sans précédent pour plusieurs raisons. Les échantillons analysés sont constitués de 3.860 ARNc, en lieu et place d'un extrait cellulaire. Parmi ceux-ci, 1.309 ARNc individuels ont été utilisés en concentration différentes dans les deux séries de données comparées. Le chip utilisé pour les mesures du niveau d'expression est le chip DrosGenome1, basé sur la version 1.0 de la séquence du génome de la drosophile, et comporte 14.010 sets d'oligonucléotides. Les expériences ont été réalisées sur 3 chips, tant pour les mesure servant de témoin que pour les mesures ou de l'ARNc a été ajouté.

Dans l'étude réalisée, CHOE utilise ce jeu de données pour évaluer plusieurs étapes du traitement des données, et plusieurs combinaisons de méthodes de prétraitement et d'analyse de l'expression différentielle. L'analyse réalisée lors de ce travail s'intéresse exclusivement à la détection des gènes différentiellement exprimés. En conséquence, parmi les nombreux jeux de données générés par CHOE au départ des mesures réalisées, nous avons choisi de travailler avec celui pour lequel la combinaison des méthodes de prétraitement est la plus fiable (CHOE *ET AL.*, 2005) [33].

Le jeu de données est publiquement disponible et est associé à une liste qui reprend, pour chaque *probeset*, le rapport réel entre les niveaux d'expression des deux séries de mesures. Cette liste est utilisée pour évaluer précisément les performances des méthodes.

VI.A.3. Jeux de données biologiques

VI.A.3.a. E-MEXP-231

Le jeu de données E-MEXP-231 a été décrit par Yap *et al.*, et rendu disponible sur la base de données *ArrayExpress* (EBI). Il se rapporte à l'étude d'adénocarcinomes primaires pulmonaires, et présente un grand nombre de mesures. Les 58 biopuces utilisées, de modèle HG-U133A sont fournies par Affymetrix, et comparent 49 échantillons issus de tissus pulmonaires atteints d'un adénocarcinome primaire et 9 échantillons relatifs à des tissus pulmonaires sains.

Nous avons utilisé ce jeu de données pour illustrer la relation entre le niveau d'expression mesuré et la variabilité, et pour étudier l'évolution de la variance obtenue en fonction du nombre de réplicats (la mesure de référence étant calculée sur le jeu complet), et lors de l'utilisation d'un estimateur de type « fenêtre », afin d'optimiser le choix de la taille de la fenêtre.

Le nombre relativement important de mesures effectuées a permis de répondre à plusieurs questions, telle que l'évaluation de la capacité de plusieurs méthodes à fournir les mêmes résultats sur un jeu de donnée complet et de taille réduite [149].

VI.A.3.b. E-MEXP-445

Le jeu de données E-MEXP-445 a été décrit par Bosco *et al.*, et est accessible sur la base de données publique *ArrayExpress*. Le modèle de micropuce utilisé est HG-U133A, fourni par Affymetrix. Le design de l'expérience repose sur la comparaison d'échantillons de monocytes humains. Parmi les échantillons d'ARN hybridés sur les 6 *microarrays* utilisées, 3 sont extraits d'une culture réalisée dans des conditions normoxiques, et les 3 autres sont extraits de cultures réalisées en conditions hypoxiques. La compréhension des mécanismes déclenchés par le manque d'oxygène est d'une grande importance pour la compréhension des voies métaboliques impliquées dans différents types de cancer, implicitement liés à la réponse hypoxique (les cellules situées au coeur d'une tumeur manquent d'oxygène) [22].

VI.A.3.c. GSE-1056

VENGELLUR *ET AL.*, en 2005, étudient la réponse hypoxique de la lignée cellulaire Hep3b (hépatocytes), et de 3 traitements connus pour simuler les conditions hypoxiques: le chlorure de cobalt (100 µM), le chlorure de nickel (100 µM), et le DFO (100 µM). L'ARNm a été extrait séparément sur deux réplicats biologiques. L'ARNc a été hybridé sur des puces Affymetrix de type *GeneChip* HG-U95Av1. Les auteurs, dans le cadre de leur étude, ont prétraité les données avec *GCRMA* et analysé l'expression différentielle sur base d'une procédure *ANOVA* ($p<0.05$), et du *fold-change* ($FC>2$). Les données sont disponibles sur *GEO*, avec l'identifiant GSE1056. Les auteurs montrent que les « simulateurs » d'hypoxie fournissent un profil d'expression différent du profil observé dans les conditions hypoxiques. Nous avons utilisé ce jeu de données (non prétraité) pour comparer les conditions normoxiques et hypoxiques. Nous lui avons appliqué le prétraitement *GCRMA*, par soucis d'uniformisation de la procédure analytique sur plusieurs jeux de données [139].

VI.A.3.d. GSE-4086

KIM *ET AL.*, en 2006, étudient la réponse hypoxique de la lignée cellulaire des lymphocytes B humains, P493-6. Deux réplicats biologiques issus de cultures différentes ont été utilisés pour extraire l'ARNm. L'ARNc a été hybridé sur des puces Affymetrix de type HG-U133A. Les auteurs ont ensuite prétraité les données avec RMA, et les ont analysées sur base du *fold-change*. Le jeu de données est disponible sur *GEO*, avec l'identifiant GSE-4086 [86]. Nous avons utilisé ce jeu de donnée, en le prétraitant avec *GCRMA*, pour étudier l'effet de la privation d'oxygène, par comparaison avec deux autres jeux de données biologiques relatifs à l'hypoxie, E-MEXP-445 (monocytes) et GSE-1056 (hépatocytes).

VI.A.3.e. E-GEOD-7429

XIAO *ET AL.*, en 2008, rapportent une étude relative à l'ostéoporose, grâce à la technologie des micoarrays. Le jeu de donnée rendu public par les auteurs compare 10 échantillons relatifs à des cas dont la densité minérale des os (BMD : *Bone Mineral Density*) est faible à 10 échantillons relatifs à des cas de BMD élevée. Les échantillons ont été prélevés sur 20 femmes blanches ménopausées depuis au moins 12 moins, agées de 54 à 60 ans, dont 10 présentent une faible BMD, et 10 présentent une BMD élevée. Leurs lymphocytes B ont

été isolés au départ de 70 ml de sang, et l'ARNm extrait a été recopié (ARNc). L'ARNc a ensuite été hybridé sur des puces Affymetrix de type GeneChip HG-133A. Les gènes différentiellement exprimés mis en évidence par les analyses ont été validés par *real-time PCR*, en deux étapes, dont la seconde est quantitative. Les analyses originales ont été réalisées sur base du prétraitement *RMA*, et analysé avec le test du *t* de STUDENT. Le jeu de donnée est disponible sur *ArrayExpress* (identifiant E-GEOD-7429), et sur *GEO* (identifiant GSE-7429) [148]. Nous avons utilisé ces données, prétraitées avec *GCRMA*, pour illustrer le développement de la méthode *FAERI*.

VI.B.
Méthodes & Procedures

VI.B.1. La méthode *window*

VI.B.1.a. Estimation de la variance sur base d'une fenêtre

Une fonction d'estimation de la variance sur base d'une fenêtre a été implémentée, en temps que composante des méthodes *window t-test* et *regularized t-test*. La procédure utilisée pour calculer la variance est schématisée dans la Figure IV.A.6 (page 117). Dans un premier temps, le niveau moyen d'expression est calculé individuellement à partir des données d'expressions disponibles. Le jeu de données est ensuite trié sur base de la moyenne calculée. Pour chaque gène, une taille de fenêtre est définie par l'utilisateur, et le nombre correspondant de gènes est utilisé de part et d'autre du gène d'intérêt au cours de l'estimation de la variance qui lui est associée. L'étape suivante consiste à calculer la somme des carrés des écarts, et des degrés de libertés, associés à chacune des séries de données composant la fenêtre. La somme des carrés des écarts totale est alors utilisée, en combinaison avec le nombre de degrés de libertés global lié à la fenêtre, pour fournir la variance associée au gène d'intérêt. La table VI.B.1 présente un tableau comparatif des procédures utilisées dans les méthodes classiques et dans les méthodes « fenêtre ».

L'algorithme utilisé pour calculer la variance sur base d'une fenêtre est le suivant [18] :

- 1 - Calculer le niveau moyen d'expression pour chaque gène et tri du jeu de données sur base de cet estimateur.
- 2 - Pour chaque gène, définir une liste de gènes situé dans une fenêtre de taille k autour du gène intérêt. Les gènes compris entre les positions ⌈n-k⌉ et ⌈n+k⌉ sont utilisés, avec n=indice du gène d'intérêt et k = le nombre de gènes à sélectionner de part et d'autre du gène d'intérêt (taille). Une fenêtre de taille k implique donc l'utilisation de ⌈2k+1⌉ gènes.
- 3 - Calculer la somme des carrés des écarts individuels et les degrés de libertés pour toutes les gènes.
- 4 - Pour caractériser la fenêtre définie (une par gène), additionner les sommes des carrés des écarts des gènes présents dans la fenêtre (SSE_{tot}), et additionner les degrés de libertés associés à chaque gène.

$t_g = \frac{R_g}{Y_g}$ $SSE = \Sigma_{i=1}^{n}(x_i - M)^2$	Student t-test Equal Variances $\sigma_1^2 = \sigma_2^2$	Welch t-test Unequal Variances $\sigma_1^2 \neq \sigma_2^2$
$S_g = \sqrt{\frac{\Sigma_c SSE_c}{\Sigma_c df_c}}$	$Y_g = S_g\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$Y_g = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
	$S_g = \sqrt{\frac{SSE_1 + SSE_2}{n_1+n_2-2}}$	$S_1 = \sqrt{\frac{SSE_1}{n_1-1}}$
	$d.f.(t) = n_1+n_2-2$	$d.f.(t) = \frac{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$

$SSE = \Sigma_{i=1}^{n}(x_i - M)^2$	Window t-test Equal Variances $\sigma_1^2 = \sigma_2^2$	Window Welch t-test Unequal Variances $\sigma_1^2 \neq \sigma_2^2$
$S_g = \sqrt{\frac{\Sigma_c SSE_c}{\Sigma_c df_c}}$	$Y_g = S_g\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$	$Y_g = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$
	$S_g = \sqrt{\frac{SSE_{0,1} + SSE_{0,2}}{N_1 - G_{0,1} + N_2 - G_2}}$	$S_1 = \sqrt{\frac{SSE_{0,1}}{N_1 - G_1}}$
$S_{0,1} = \sqrt{\frac{SSE_{0,1}}{N_1 - G_1}}$ $SSE_{0,1} = \Sigma_{p=1}^{G} SSE_{1,p}$	$d.f.(t) = n_1+n_2-2$	$d.f.(t) = \frac{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$

Table VI.B.1

Comparaison des procédures d'analyse classiques (STUDENT *t-test* et WELCH *t-test*) avec leurs équivalentes basées sur l'utilisation de la relation empirique entre le niveau d'expression moyen et la variabilité (*window t-test* et *window* WELCH *t-test*).

- ☞ 5 - Pour estimer la variance associée à une fenêtre, calculer le rapport entre la somme des carrés des écarts totale (SSE_{tot}) et le total des degrés de liberté.
- ☞ 6 - Répéter les opérations 4 à 5 pour chaque gène.

VI.B.1.b. Etude de l'influence de la taille de la fenêtre en fonction du nombre de mesures.

Le nombre de *probesets* inclus dans la fenêtre utilisée pour estimer la variance est un paramètre très important dans le contexte analytique envisagé. Deux situations extrêmes peuvent se présenter: une fenêtre qui se défini exclusivement par le *probeset* d'intérêt mène à une estimation de la variance individuelle limitée par le nombre de mesures. A l'inverse, une fenêtre de taille importante mène à une constante qui ne reflète aucune variabilité individuelle, qui est d'autant mieux estimée que le nombre de *probesets* considérés est important et que le nombre de mesures répétées est élevé. Entre ces deux extrêmes, il existe un grand nombre de situations intermédiaires, pour lesquelles le poids du *probeset* d'intérêt est dilué par le nombre de *probesets* inclus dans la fenêtre. La taille d'utilisation optimale d'une fenêtre est donc une fonction du nombre de réplicats disponibles [18]. Pour étudier l'évolution de l'estimation de la variance au départ d'une fenêtre en fonction de sa taille et du nombre de mesures, l'algorithme suivant a été utilisé sur le jeu de données E-MEXP-231, qui présente 49 mesures relatives à une même condition expérimentale :

- ☞ 1 - Pour chaque *probeset*, calculer l'écart-type en utilisant la totalité des mesures disponibles (sd_{ref}) ;
- ☞ 2 - Pour un taille de fenêtre donnée, calculer l'écart-type en utilisant l'ensemble des *probesets* inclus dans la fenêtre, pour chaque *probeset*, en suivant la procédure décrite au paragraphe IV.A.2.b. p. 108. (sd_W) ;
- ☞ 3 - Répéter l'étape 2 pour un ensemble de jeux de données de taille restreinte, générés par échantillonage aléatoire au sein des 49 mesures disponibles. Les études rapportées dans ce travail font intervenir respectivement 2, 3, 4, 5, 6, 8, 10, 15 et 20 réplicats.
- ☞ 4 - Répéter les étapes 2 et 3 pour différentes tailles de fenêtres. La gamme testée dans ce travail s'étend d'une taille de fenêtre de 0 (estimateur classique) à 30 (estimateur fenêtre utilisant 61 *probesets*).

Pour évaluer les conditions optimales d'utilisation d'une fenêtre, nous avons défini un

indicateur global de la qualité de l'estimation, par comparaison avec la valeur attendue, définie dans l'étape 1 sur base du jeu complet:

- ☞ 5 - Pour chaque estimateur calculé aux étapes 1 à 4, calculer l'erreur relative (R.E.), définie par la formule VI.B.1.

$$R.E.(i) = \frac{|sd_{W,i} - sd_{ref,i}|}{sd_{ref,i}} \quad \text{(Equ. VI.B.1)}$$

- ☞ 6 - Pour un même jeu de paramètres (taille de fenêtre et nombre de mesures), calculer la médiane de l'erreur relative de l'ensemble des *probesets* représentés ;

- ☞ 7 - Répéter l'étape 6 pour chaque jeu de paramètres.

Les valeurs obtenues au terme de l'étape 7 peuvent ensuite être portée en graphique, en reliant les points relatifs à un même nombre de mesures, de façon à obtenir un ensemble de courbes représentant l'évolution de l'erreur commise lors de l'estimation en fonction de la taille la fenêtre.

> ✍ A titre complémentaire, la dispersité de l'erreur commise a également été suivie en calculant la déviation absolue (MAD), au cours de l'étape 6, au lieu de la médiane.

VI.B.2. La méthode *consensus*

L'évaluation du *consensus* des résultats a pour objectif de fournir une estimation robuste de l'expression différentielle, au départ de plusieurs méthodes. Statistiquement parlant, la probabilité qu'un *probeset* non impliqué (hypothèse nulle) soit détecté par plusieurs méthodes se calcule par le produit des probabilités obtenues par chaque méthode. La *p-value* du *consensus* peut donc être calculé par le produit des *p-values* obtenues pour chaque méthode (Equation VI.B.2).

$$p_{cons}(i) = \prod_{k=1}^{n_{meth}} p_{ik} \quad \text{(Equ. VI.B.2)}$$

Néanmoins, la comparaison du *consensus* obtenu avec un nombre variable de méthodes révèle une dépendance du niveau de significativité atteint et du nombre de méthodes utilisées. Afin de permettre une comparaison croisée des résultats du *consensus* lorsque le nombre de méthodes sélectionnées est différent, nous avons adapté la procédure mathématique envisagée. L'équation VI.B.2 peut être reformulée sur base d'opérations logarithmiques, où le logarithme du produit des probabilités se traduit en une somme des logarithmes des probabilités méthodes-spécifiques (Equation VI.B.3). Pour rendre le résultats indépendants du nombre de méthodes utilisées, nous avons ajusté l'évaluation du *consensus* en remplaçant la somme des logarithmes par leur moyenne, tel que le montre l'équation VI.B.4.

$$\log\left(p_{cons}(i)\right) = \sum_{k=1}^{n_{meth}} \log\left(p_{ik}\right) \quad \text{(Equ. VI.B.3)}$$

$$\log\left(cons.score(i)\right) = \frac{\sum_{k=1}^{n_{meth}} \log\left(p_{ik}\right)}{n_{meth}} \quad \text{(Equ. VI.B.4)}$$

Le deuxième point important rencontré lors de l'évaluation du *consensus* sur base des probabilités individuelles est lié à la distribution des *p-values* des différentes méthodes. Chaque méthodologie est caractérisée en partie par la manière dont la significativité est évaluée, et sur l'impact des corrections utilisées par leur procédure. Il en découle une liste de *p-values* dont l'échelle/la distribution est variable entre les différentes méthodes, y compris lorsque les résultats sont proches. La conséquence de cette variabilité distributionnelle implique que les méthodes associées à une gamme plus large de *p-values*

prennent un poids plus important lors de l'évaluation du *consensus*, si bien qu'il faut s'assurer au préalable d'uniformiser la distribution des *p-values*. Plusieurs études s'intéressent à cette problématique par le biais d'une correction additionnelle, basée sur le FDR (*False Discovery Rate*). A notre connaissance, cette approche ne solutionne pas complètement le problème distributionnel soulevé. Alternativement, le *consensus* peut être appliqué sur les rangs des résultats en lieu et place de leurs *p-values* (en divisant par le nombre total de *probesets* pour rester dans une gamme de valeurs comprises entre 0 et 1). Ainsi, quelle que soit la méthode utilisée, la substitution des *p-values* par leur rang fourni une distribution commune. En revanche, le *consensus* évalué au départ des rangs ne constitue pas une probabilité, mais fourni un score caractéristique de la position du *probeset* dans la liste des résultats (distribué entre 0 et 1). La séquence finale des scores évalués constitue donc la séquence attendue des *probesets* relative à l'expression différentielle.

Enfin, l'utilisation de différentes méthodes d'analyse montre que certaines méthodes sont plus fiables que d'autres, tel que décrit dans la première partie des résultats présentés dans ce travail. Il importe donc de permettre la prise en compte des performances lors de l'évaluation du *consensus*, afin de donner plus de poids aux meilleures méthodes. A titre d'exemple, l'utilisation d'un nombre réduit de mesures fourni de meilleurs résultats avec des méthodes de correction de la variance (*regularized t-test*, *window t-test*, *Limma*, *shrinkage t*) qu'avec le test de STUDENT classique. De même, les procédures ayant recours à la correction de WELCH sont plus appropriés lorsque les variances individuelles sont différentes entre les échantillons comparés [143].

L'équation VI.B.5 présente la formulation générale du calcul du score *consensus*, « sachant que des méthodes particulières fournissent de meilleurs résultats sur base de propriétés intrinsèques du jeux de données ». La moyenne des logarithmes utilisée précédemment y est remplacée par une moyenne pondérée.

$$\log\left(cons.score(i)\right) = \frac{\sum_{k=1}^{n_{meth}} w_k \log\left(score_{ik}\right)}{\sum_{k=1}^{n_{meth}} w_k} \quad \text{(Equ. VI.B.5)}$$

où $score_{ik}$ corresponds a la *p-value* individuelle associée au gène i ou à son rang par la méthode k.

VI.B.3. La méthode FAERI

VI.B.3.a. Optimisation du calcul de la somme des carrés des écarts

L'estimation de la somme des carrés des écarts au départ des données, au sein de la procédure *ANOVA* et de la méthode *FAERI*, peut être reformulée afin de réduire le nombre d'opérations nécessaires, et ainsi réduire le temps de calcul. L'équation VI.B.6 présente les opérations mathématiques qui conduisent à une formulation simplifiée du calcul de la somme des carrés des écarts.

$$
\begin{aligned}
&\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 \\
&=\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2X_i\bar{X}\right) \\
&=\sum_{i=1}^{n}\left(X_i^2\right)+n\,\bar{X}^2-2\,\bar{X}\sum_{i=0}^{n}X_i \\
&=\sum_{i=1}^{n}\left(X_i^2\right)+n\,\bar{X}^2-2\mathrm{n}\,\bar{X}\,\bar{X} \qquad \text{(Equ. VI.B.6)} \\
&=\sum_{i=1}^{n}\left(X_i^2\right)-n\,\bar{X}^2 \\
&=\sum_{i=1}^{n}\left(X_i^2\right)-\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}
\end{aligned}
$$

Sur base de cette équation simplifiée, le calcul de la somme des carrés des écarts repose uniquement sur une matrice de données d'expressions, et sur le carré de celle-ci [55, 56, 119].

VI.B.3.b. Calcul de la statistique F caractéristique d'un groupe de gènes

L'algorithme de calcul de calcul de la statistique F associé à la méthode *FAERI* est identique à la procédure utilisée dans un test *ANOVA* à 2 facteurs croisés. La procédure optimisée d'évaluation de la statistique de groupe de gènes n'est valable que pour un nombre de mesures identiques entre les conditions comparées ($n_1=n_2$), et repose sur les étapes suivantes [55, 56, 119] :

☞ 1 - Calculer les sommes des carrés des écarts associés à l'expérience, aux gènes, à l'erreur résiduelle, à l'interaction entre les facteurs expérience et gène, et de la somme des carrés des écarts totales. En représentant la matrice des données d'expressions par X et en tenant compte de l'algorithme présenté dans l'équation VI.B.6, les carrés des écarts se calculent suivant les équations VI.B.7 à VI.B.13, où a et b représentent le nombre de niveaux des facteurs A et B, respectivement, et n_{ij} représente le nombre de mesures répétées. T_A^2 symbolise le vecteur des totaux au carré de chaque niveau de A. T_B^2 et T_{ij}^2 symbolisent les mêmes statistiques, associées au facteur B et au cellules, respectivement. FC symbolise le facteur de correction.

$$N = a\,b\,n_{ij}$$

$$FC = \frac{\left(\sum X\right)^2}{N} \quad \text{(Equ. VI.B.7)}$$

$$SS_{tot} = \sum X^2 - FC \quad \text{(Equ. VI.B.8)}$$

$$SS_A = \frac{\left(\sum T_A^2\right)}{b\,n_{ij}} - FC \quad \text{(Equ. VI.B.9)}$$

$$SS_B = \frac{\left(\sum T_B^2\right)}{a\,n_{ij}} - FC \quad \text{(Equ. VI.B.10)}$$

$$SS_{cell} = \frac{\sum T_{ij}^2}{n_{ij}} - FC \quad \text{(Equ. VI.B.11)}$$

$$SS_{AB} = SS_{cell} - SS_A - SS_B \quad \text{(Equ. VI.B.12)}$$

$$SS_{error} = SS_{tot} - SS_{cell} \quad \text{(Equ. VI.B.13)}$$

☞ 2 - Calculer les degrés de libertés associés, sur base des formules VI.B.14 à VI.B.18.

$$df_{tot} = N - 1 \quad \text{(Equ. VI.B.14)}$$

$$df_A = a - 1 \quad \text{(Equ. VI.B.15)}$$

$$df_B = b - 1 \quad \text{(Equ. VI.B.16)}$$

$$df_{AB} = (a-1)(b-1) \quad \text{(Equ. VI.B.17)}$$

$$df_{error} = a\,b(n-1) \quad \text{(Equ. VI.B.18)}$$

☞ 3 - Calculer les carrés moyens sur base de la somme des carrés des écarts et des degrés de libertés, pour chaque critère évalué (Equations VI.B.19 à VI.B.22).

$$MS_A = \frac{SS_A}{df_A} \quad \text{(Equ. VI.B.19)}$$

$$MS_B = \frac{SS_B}{df_B} \quad \text{(Equ. VI.B.20)}$$

$$MS_{AB} = \frac{SS_{AB}}{df_{AB}} \quad \text{(Equ. VI.B.21)}$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} \quad \text{(Equ. VI.B.22)}$$

☞ 4 - Calculer la statistique *F* associée à chaque critère étudié, sur base des équations VI.B.23 à VI.B.25.

$$F_A = \frac{MS_A}{MS_{error}} \quad \text{(Equ. VI.B.23)}$$

$$F_B = \frac{MS_B}{MS_{error}} \quad \text{(Equ. VI.B.24)}$$

$$F_{AB} = \frac{MS_{AB}}{MS_{error}} \quad \text{(Equ. VI.B.25)}$$

Cette procédure est utilisée en accord avec le modèle d'analyse de la variance à deux critères de classification croisés, où les deux critères sont considérés comme fixes. Dans le contexte de *l'ANOVA*, les différentes statistiques *F* sont comparées aux distributions théoriques, alors que dans le cas de la méthode *FAERI*, le non respect des conditions d'application de *l'ANOVA* impose le calcul d'une distribution nulle [55, 56, 119].

VI.B.4. Evaluation des performances & simulations

VI.B.4.a. Procédure de calcul des indicateurs de performances

L'algorithme utilisé pour calculer les coordonnées des figures de performances s'appuie sur la liste complète des *p-values*, pour chaque méthode. Le point de départ de la procédure consiste a utiliser la plus petite valeur de *p-value* disponible sur l'ensemble des méthodes, car l'échelle des valeurs fournies diffère entre les méthodes. L'évaluation des performances des meilleures résultats caractérise les performances associées aux *p-values* les plus petites, qui concernent les *probesets* les plus informatifs. En conséquence, la procédure d'évaluation repose sur une progression régulière au sein de la liste des logarithmes des *p-values*, pour obtenir une résolution accrue lorsque le résultat des analyse est proche de 0 (significatif). L'algorithme final est le suivant:

- 1 - Définir l'origine de la procédure en utilisant le minimum des *p-values* calculées (min.pval) .
- 2 - Calculer le logarithme en base 10 de la valeur minimale (log.min.pval) .
- 3 - Calculer les valeurs définissant 1000 intervalles de taille constante (log.int) entre log.min.pval et 0 (associé à une *p-value* = 1);
- 4 - Transposer la définition des intervalles à la liste des *p-values*, en calculant int = $10^{log.int}$.
- 5 - Pour chaque intervalle défini, calculer le FDR, la sensibilité, la spécificité, sur base du nombre de vrais et faux positifs et négatifs (établis grâce à une connaissance des résultats attendus), pour chaque méthodologie.

VI.B.4.b. Simulation de données aléatoires

Plusieurs approches ont été précédemment décrite pour simuler des données issues de microarrays. Nous avons choisi d'utiliser le modèle proposé par Shaik & Yeasin en 2007, au départ d'une distribution normale [123]. Au sein d'un jeu de données biologiques, les gènes représentés peuvent être exprimés identiquement ou différentiellement, quelque soit le niveau d'expression atteint. Le scénario envisagé par les auteurs ne tient pas compte du

niveau d'expression variable des gènes. Nous avons donc adapté leur procédure, en simulant des gènes répartis dans ces différentes catégories :

- ☞ 9.600 gènes non exprimés, simulés autour d'une moyenne égale à 0 ;
- ☞ 1.000 gènes exprimés à 5 différents niveaux d'expression différents (5000 au total), autour de moyennes égale à 1, 2, 3, 4 et 5 ;
- ☞ 20 gènes différentiellement exprimés, dont 10 gènes activés et 10 gènes réprimés, pour 10 niveaux de différences d'expression (200 au total), simulés autour d'une moyenne égale à 0 pour la moitié des mesures, et égale à chaque niveau d'expression successivement pour la seconde moitié des mesures ;
- ☞ dans tous les cas, la variance a été simulée sur base d'une distribution gamma de moyenne = 2 et variance = 2 [123].

Le jeu de donnée final comporte donc 15.200 gènes simulés, dont 200 sont différentiellement exprimés, avec 3 réplicats pour chaque condition simulée. La simulation réalisée a été reproduite 1000 fois, et les résultats sous forme de p-values ont été rassemblés, pour être évalués en une seule étape, sur base de la connaissance précise des différences simulées.

VI.B.4.c. Simulation de données et de groupes aléatoires

Sur base de données simulées, plusieurs groupes de gènes peuvent être définis de façon à mettre en évidence les critères influant sur les performances, et les faiblesses des différentes méthodes. Reproduire parfaitement la structure d'un jeu de données biologique réel est une utopie à l'heure ou les relations entre les gènes ne sont pas totalement répertoriées. Néanmoins, un tel scénario de simulation permet d'étudier globalement les performances des méthodes d'analyse, sur base de la transposition de critères connus et observés dans les jeux de données publics.

Les simulations aléatoires réalisées en analyse de groupes de gènes dérivent du scénario décrit ci-dessus pour l'évaluation de l'analyse individuelle. Les données ont été simulées cette fois avec 10 niveaux de différences d'expression, sur le même scénario. 40 gènes ont été sélectionnés pour chacun de ces niveaux, dont 50% en sur-expression et 50% en sous-expression. Au total 400 gènes différentiellement exprimés ont été simulés, ainsi que 9.600 gènes non exprimés et 10.000 gènes exprimés identiquement à 10 niveaux d'expression. Le

jeu de données final comporte donc 20.000 gènes, dont 2% sont différentiellement exprimés (400 gènes).

Les groupes de gènes ont été définis aléatoirement sur base de ces gènes et des gènes identiquement exprimés. Au total, 50.000 groupes ont été générés aléatoirement, parmi lesquels 2.500 ont été sélectionnés aléatoirement parmi tous les gènes différentiellement exprimés, et 47.500 ont été générés au départ des gènes identiquement exprimés (à tous les niveaux). Les groupes générés sont soit de taille fixe (3 gènes), soit de taille aléatoire (entre 2 et 100 gènes). La simulation correspond donc à une analyse de 500 groupes, reproduite 100 fois, dont 5% (25 groupes) seraient impliqués, sans corrélation entre les membres, exprimés à des niveaux variables, et présentant une réponse variable.

VI.B.4.d. Simulation de données aléatoires corrélées

D'un point de vue biologique, l'expression des gènes est régulée de diverses manières. Le terme de coexpression est employé lorsque les mesures associées à plusieurs gènes sont corrélées. Les études de coexpression et de *clustering* visent à définir des groupes de gènes, qui peuvent ensuite être utilisés, notamment à des fins diagnostique. La définition de ces groupes s'éloigne donc de la stratégie de simulation évoquée au paragraphe précédent, où chaque gène est généré indépendamment des autres.

Pour simuler les différents cas de figure possibles, nous avons choisi de suivre la stratégie adoptée par Marit Ackermann, qui repose sur la simulation de 9 types de groupes. Les critères définis pour les constituer sont le nombre de gènes différentiellement exprimés au sein du groupe, la direction de la différence, et la corrélation. Pour chaque type de groupe défini, la table VI.B.2 présente les paramètres utilisés lors de la simulation. Chaque groupe a été simulé 100 fois.

La genèse de données corrélées repose sur la définition d'un matrice de covariance (ou de corrélation). Celle-ci est utilisée pour générée des données aléatoires. Ces opérations sont réalisées grâce à la fonction *mvrnorm (R-package MASS)*, sur base des scripts écrits en R par Marit Ackermann (Marit Ackermann & Korbinian Strimmer, communication personnelle).

	Différence d'expression	Corrélation	Diff. Exprimés	Sur-exprimés	Sous-exprimés	Design
Set n°1	0.75	0.6	20	20	0	uni
Set n°2	0.75	0	20	20	0	uni
Set n°3	0	0	0	0	0	
Set n°4	0.75	0.6	10	10	0	uni
Set n°5	0.75	0	10	10	0	uni
Set n°6	1	0.6	20	10	10	bidir
Set n°7	1	0	20	10	10	bidir
Set n°8	1	0.6	10	5	5	bidir
Set n°9	1	0	10	5	5	bidir

Table VI.B.2
Définition des séries de mesures utilisées pour évaluer les performances des méthodes d'analyse de l'expression différentielle de groupes de gènes.

VI.B.4.e. Création d'un jeu de données « réel » d'évaluation

Afin de réaliser une évaluation des performances la plus représentative de la réalité biologique possible, une méthode de *benchmark* originale imaginée dans le cadre de ce travail a été mise en oeuvre par BERTRAND DE MEULDER, sous la coordination de BENOÎT DE HERTOGH.

La procédure repose sur une collection de jeux de données authentiques issus du domaine public et leur concaténation dans une matrice unique comportant 1.292.314 séries de (2*15) mesures. Cette matrice est nommée la matrice totale. Celle-ci est ensuite triée en fonction du rapport D/S, dont les paramètres sont définis par les équations VI.B.26 et VI.B.27.

$$D=\mu_2-\mu_1 \quad \text{(Equ. VI.B.26)}$$

$$S=\frac{S_1+S_2}{2} \quad \text{(Equ. VI.B.27)}$$

où D est la différence entre les moyennes associées aux deux conditions comparées, μ_1 et μ_2, et S est la moyenne des déviations standard mesurées pour les deux conditions.

La matrice est ensuite triée de sorte que les *probesets* dont le rapport D/S est maximal soient positionnés dans les premières lignes de la matrice.

Une valeur seuil du rapport D/S est ensuite évaluée sur base de l'équation VI.B.28.

$$PPP = \frac{(1-\beta)P}{(1-\beta)P + \alpha(1-P)} \quad \text{(Equ. VI.B.28)}$$

où PPP est le pouvoir prédictif positif, et P est la prévalence.

L'expression de α en fonction de β, P et PPP, conduit à l'expression VI.B.29.

$$\alpha = \frac{\frac{(1-\beta)P}{PPP} - (1-\beta)P}{1-P} \quad \text{(Equ. VI.B.29)}$$

Sur base de l'équation VI.B.30, en fixant $D = M_2 - M_1$ pour estimer $\mu_2 - \mu_1$, de variance $2\sigma^2$ et en estimant σ sur base de la valeur de S, nous obtenons l'équation VI.B.31.

$$n \geq \frac{\left(z_{1-\alpha} - z_{1-\beta}\right)^2 \sigma^2}{\left(\mu_0 - \mu_1\right)^2} \quad \text{(Equ. VI.B.30)}$$

$$\left(D/S\right)_{cutoff} = \frac{\sqrt{2}\left|z_{1-\alpha} + z_{1-\beta}\right|}{\sqrt{n}} \quad \text{(Equ. VI.B.31)}$$

La valeur seuil du rapport D/S est donc obtenue en choisissant les paramètres de confiance $(1-\alpha)$, de puissance $(1-\beta)$ et le nombre de réplicats (n).

Les 200 *probesets* dont le rapport D/S est supérieur au seuil calculé sont sélectionnés et considérés comme différentiellement exprimés.

Un second seuil est fixé sur le rapport D/S, sur base d'une puissance inférieure de 10% à celle utilisée pour le premier seuil défini. Ce second seuil est utilisé pour déterminer les *probesets* identiquement exprimés (le *background*).

Grâce aux deux seuils fixés, une nouvelle matrice est définie, à partir de la matrice totale, en choisissant les 200 *probesets* définis comme différentiellement exprimés, et 19.800 *probesets* identiquement exprimés (sélection aléatoire). Pour chaque *probeset*, n réplicats sont sélectionnés aléatoirement parmi les 15 réplicats disponibles pour chaque condition.

La collection de jeux de données de taille réduite générées par ce principe est ensuite utilisée avec le logiciel *PEGASE* pour effectuer l'analyse statistique et l'évaluation des performances, sachant que les 200 premiers *probesets* constituent la « vérité ».

VII. Références

1. Ackermann M. & Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics.*, 2009, 10:47. doi:10.1186/1471-2105-10-47

2. Affymetrix. Technical note: guide to probe logarithmic intensity error (PLIER) estimation. 2005. *http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf.*

3. Affymetrix. Latin Square Data for Expression Algorithm Assessment. Human Genome U95 Data Set. *http://www.affymetrix.com/support/technical/sample_data/datasets.affx.*

4. Affymetrix. Latin Square Data for Expression Algorithm Assessment. Human Genome U133 Data Set. *http://www.affymetrix.com/support/technical/sample_data/datasets.affx.*

5. Aittokallio T., Kurki M., Nevalainen O., Nikula T., West A. & Lahesmaa R. Computational strategies for analysing data in gene expression microarray experiments. *J. Bioinform. Comput. Biol.*, 2003, 1 : 541-586.

6. Al-Shahrour F., Minguez P., Vaquerizas JM., Conde L. & Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.*, 2005, 33 (Web Server issue) : W460-4.

7. Al-Shahrour F., Díaz-Uriarte R. & Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, 2005, 21(13) : 2988-93. Epub 2005 Apr 19.

8. Allison D.B., Cui X., Page G.P. & Sabripour M. Microarray data analysis: from disarray to consolidation and *consensus. Nat. Rev. Genet.*, 2006, 7 : 55-65.

9. Altman R.B. & Raychaudhuri S. Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, 2001, 11 : 340-347.

10. Audic S. & Claverie J.M. The significance of digital gene expression profiles. *Genome Res.*, 1997, 7(10) : 986-95.

11. Baldi P. & Long A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 2001, 17 : 509-519.

12. Barrera L., Benner C., Tao Y.C., Winzeler E. & Zhou Y. Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays. *BMC Bioinformatics*, 2004, 5 : 42.

13. Barry W.T., Nobel A.B. & Wright F.A. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 2005, 21(9) : 1943-9. Epub 2005 Jan 12.

14. Bayes T. An Essay towards solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society of London*, 1763, 53 : 370-418.

15. Bayes T. Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances, *Biometrika*, 1763/1958, 45 : 296-315.

16. Benjamini Y. & Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 1995, 57 : 289-300.

17. Benjamini Y. & Yekutieli D. Quantitative trait Loci analysis using the false discovery rate. *Genetics*, 2005, 171(2) : 783-90. Epub 2005 Jun 14.

18. Berger F., De Hertogh B., Pierre M., Gaigneaux A. & Depiereux E. The "Window t-test": a simple and powerful approach to detect differentially expressed genes in microarray datasets. *Cent. Eur. J. Biol.*, 2008, 3 : 327-344.

19. Berger F., De Hertogh B., Bareke E., Pierre M., Gaigneaux A. & Depiereux E. PHOENIX: a web-interface for (re)analyses of microarray data. *Cent. Eur. J. Biol.*, 2009, 4(4) : 603 : 618.

20. Bonferroni C. E. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, 1935 Rome: Italy, 13-60.

21. Bonferroni C. E. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936, 8 : 3-62.

22. Bosco M.C., Puppo M., Santangelo C., Anfosso L., Pfeffer U., Fardin P., Battaglia F. & Varesio L. Hypoxia Modifies the Transcriptome of Primary Human Monocytes: Modulation of Novel Immune-Related Genes and Identification Of CC-Chemokine Ligand 20 as a New Hypoxia-Inducible Gene. *J. Immunol.*, 2006, 177 : 1941-1955.

23. Brazma A., Jonassen I., Vilo J. & Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, 1998, 11 : 1202-1215.

24. Breitling R., Armengaud P., Amtmann A. & Herzyk P. Rank Products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, 2004, 573(1-3) : 83-92.

25. Breitling R., Amtmann A. & Herzyk P. Graph-based iterative Group Analysis enhances microarray interpretation. *BMC Bioinformatics*, 2004, 5 : 100.

26. Breitling R., Amtmann A. & Herzyk P. Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*. 2004, 5 : 34.

27. Broberg P. Statistical methods for ranking differentially expressed genes. *Genome Biol.*, 2003, 4(6) : R41. Epub 2003 May 29.

28. Bucher P. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, 1999, 9 : 400-407.

29. Bussemaker H.J., Li H. & Siggia E.D. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA*, 2000, 97 : 10096-10100.

30. Cheadle C., Becker K.G., Cho-Chung Y.S., Nesterova M., Watkins T., Wood III W., Prabhu V. & Barnes K.C. A rapid method for microarray cross-platform comparisons using gene expression signatures. *Molecular and Cellular Probes*, 2007, 21 : 35-46. doi: 10.1016/j.mcp.2006.07.004.

31. Chen Z., McGee M., Liu Q., Kong M., Deng Y. & Scheuermann R.H. A Distribution-Free Convolution Model for background correction of oligonucleotide microarray data. *BMC Genomics*, 2009, 10 Suppl 1 : S19.

32. Cho R.J., Huang M., Campbell M.J., Dong H., Steinmetz L., Sapinoso L., Hampton G., Elledge S.J., Davis R.W. & Lockhart D.J. Transcriptional regulation and function during the human cell cycle. *Nat. Genet.* 2001, 27 : 48-54

33. Choe S.E., Boutros M., Michelson A.M., Church G.M. & Halfon M.S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control data set. *Genome Biol.*, 2005, 6 : R16.

34. Cohen B.A., Mitra R.D., Hughes J.D. & Church G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.*, 2000, 26 : 183-186.

35. Conlon E.M., Song J.J. & Liu J.S. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, 2006, 7 : 247. doi: 10.1186/1471-2105-7-247.

36. Cope L.M., Irizarry R.A., Jaffee H.A., Wu Z. & Speed T.P. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 2004, 20 : 323-331.

37. Cui X. & Churchill G.A. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, 2003, 4 : 210.

38. Cui C., Gene Hwang J.T., Qiu J., Blades N.J. & Churchill G.A. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 2005, 6 : 59-75.

39. Delmar P., Robin S. & Daudin J.J. VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. *Bioinformatics*, 2005, 21(4) : 502-8. Epub 2004 Sep 16.

40. Dempster A.P. A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 1958, 29 : 995-1010.

41. Dempster A.P. A significance test for the separation of two highly multivariate small samples. *Biometrics*, 1960, 16 : 41-50.

42. De Hertogh B., De Meulder B., Berger F., Pierre M., Bareke E., Gaigneaux A. & Depiereux E. Benchmark of microarray data preserving the actual variance. *Bioinformatics, under review.*

43. De Risi J.L., Iyer V.R. & Brown P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997, 278, 680-686.

44. Dinu I., Potter J.D., Mueller T., Liu Q., Adewale A.J., Jhangri G.S., Einecke G., Famulski K.S., Halloran P. & Yasui Y. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 2007, 8 : 242.

45. Draghici S., Khatri P., Bhavsar P., Shah A., Krawetz S.A. & Tainsky M.A. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, 2003, 31(13) : 3775-81.

46. Dudoit S., Yang Y.H., Callow M.J. & Speed T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 2002, 12 : 111-139.

47. Edgar R., Domrachev M. & Lash A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, 2002, 30 : 207-210.

48. Edelman E., Porello A., Guinney J., Balakumaran B., Bild A., Febbo P.G. & Mukherjee S. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual sample in genome-wide expression profiles. *Bioinformatics*, 2006, 22(14) : e108-e116. doi: 10.1093/bioinformatics/btl231.

49. Efron B., Tibshirani R., Storey J. & Tusher V. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 2001, 96 : 1151-1160.

50. Efron B., Storey J. & Tibshirani R. Microarrays, empirical Bayes methods, and false discovery rates. *Technical report* (2001), Stanford Univ., *http://wwwstat.stanford.edu/tibs/research.html.*

51. Efron B. & Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.*, 2002, 23(1) : 70-86.

52. Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.*, 2004, 99 : 96–104.

53. Efron B. & Tibshirani R. On testing the significance of sets of genes. *Ann. Appl. Stat.*, *2007*, 1(1) : 107-129.

54. Farmer P., Bonnefoi H. Becette V., Tubiana-Hukin M., Fumoleau P., Larsimont D., MacGrogan G., Bergh J., Cameron D., Goldstein D., Duss S., Nicoulaz A.L., Brisken C., Fiche M., Delorenzi M. & Iggo R. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, 2005, 24 : 4660-4671.

55. Feytmans E. Statistiques élémentaires. *Syllabus du cours donné aux étudiants de première année en sciences, pharmacie, médecine et médecine vétérinaire*, 163–174.

56. Feytmans E. Biostatistiques. *Syllabus du cours de biostatistiques*, 65–70.

57. Fisher, L.D. & van Belle, G. Biostatistics: A Methodology for the Health Sciences. *John Wiley and Sons*, New York (1993).

58. Fox R.J. & Dimmic M.W. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*. 2006, 7 : 126.

59. Fujita, A., Sato, J.R., de Oliveira Rodrigues, L., Ferreira, C.E., & Sogayar, M.C. Evaluating different methods of microarray data normalization. *Bioinformatics*, 2006, 7 : 469.

60. Gaigneaux A., De Hertogh B., Berger F., Pierre M., Bareke E. & Depiereux E. Discussion about ROC curves and others figures used to compare microarray statistical analyses. *Proceedings of Benelux Bioinformatics Conference* (11-13 december 2008, Maastricht, Holland), 2008.

61. Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.Y. & Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 2004, *5* : R80.

62. Giles P.J. & Kipling D. Normality of oligonucleotide microarray data and implications for parametric statistical analysis. *Bioinformatics*, 2003, 19(17) : 2254-2262. doi:10.1093/bioinformatics/btg311.

63. Goeman J.J., Geer S.A., de Kort F. & van Houwelingen H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004, 20 : 93-99.

64. Goeman J.J. & Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 2007, *23*(8) : 980-7. Epub 2007 Feb 15.

65. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D. & Lander E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 1999, 286 : 531-537.

66. Harr B. & Schlötterer C. Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic Acids Res.*, 2006, 34, *http://nar.oxfordjournals.org/cgi/content/abstract/34/2/e8*

67. Hatfield G.W., Hung S.P. & Baldi P. Differential analysis of DNA microarray gene expression data. *Mol Microbiol.*, 2003, 47(4) : 871-7. Review.

68. Hong F., Breitling R., McEntee C.W., Wittner B.S., Nemhauser J.L. & Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 2006, 22(22) : 2825-7. Epub 2006 Sep 18.

69. Hosack D.A., Dennis Jr G., Sherman B.T., Lane H.C. & Lempicki R.A. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 2003, 4 : R70.

70. Hotelling H. *Ann. Math. Statist.*, 1931, 2(3) : 360-378.

71. Hubbell E., Liu W.M. & Mei R. Robust estimators for expression analysis. *Bioinformatics*, 2002, 18 : 1585-1592.

72. Huber B.R. & Bulyk M.L. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinformatics*, 2006, 7 : 229. doi:10.1186/1471-2105-7-229.

73. Hughes J.D., Estep P.W., Tavazoie S. & Church G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, 2000, 296 : 1205-1214.

74. Hummel M. , Meister R. & Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, 2008, 24(1) : 78-85.

75. Hung S.P., Baldi P. & Hatfield G.W. Global gene expression profiling in Escherichia coli K12. The effects of leucine-responsive regulatory protein. *J Biol Chem.*, 2002, 277(43) : 40309-23. Epub 2002 Jul 18.

76. Irizarry R.A., Bolstad B.M., Collin F., Cope L.M., Hobbs B. & Speed T.P. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, 2003, 31 : e15.

77. Jain N., Thatte J., Braciale T., Ley K., O'Connell M. & Lee J.K. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, 2003, 19 : 1945-1951.

78. Jeffery I.B., Higgins D.G. & Culhane A.C. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. *BMC Bioinformatics*, 2006, 7 : 359. doi: 10.1186/1471-2105-7-359.

79. Kal A.J., van Zonneveld A.J., Benes V., van den Berg M., Koerkamp M.G., Albermann K., Strack N., Ruijter J.M., Richter A., Dujon B., Ansorge W. & Tabak H.F. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell.*, 1999, 10(6) : 1859-72.

80. Keller A., Backes C. & Lenhof H.P. Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinformatics.*, 2007, 8 : 290.

81. Kerr M.K. & Churchill G.A. Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, 2001, 77 : 123-128.

82. Khatri P., Draghici S., Ostermeier G.C. & Krawetz S.A. Profiling gene expression using onto-express. *Genomics.*, 2002, 79(2) : 266-70.

83. Khatri P. & Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, 2005, 21(18) : 3587-95. Epub 2005 Jun 30.

84. Khatri P., Sellamuthu S., Malhotra P., Amin K., Done A. & Draghici S. Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res.*, 2005, 33(Web Server issue) : W762-5.

85. Khatri P., Done B., Rao A., Done A. & Draghici S. A semantic analysis of the annotations of the human genome. *Bioinformatics.* 2005, 21(16) : 3416-21. Epub 2005 Jun 14.

86. Kim J.W., Tchernyshyov I., Semenza G.L. & Dang C.V. HIF-1-mediated expression of pyruvate dehydrogenase kinase: A metabolic switch required for cellular adaptation to hypoxia. *Cell*

Metabolism, 2006, 3 : 177-185. doi:10.1016/j.cmet.2006.02.002.

87. Kim S.Y. & Volsky D.J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*. 2005, 6 : 144.

88. Kim S.-Y. & Kim Y.S. Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data. *BMC Bioinformatics*, 2006, 7 : 330. doi:10.1186/1471-2105-7-330.

89. Kim S.-Y. & Kim Y.S. A gene sets approach for identifying prognostic gene signatures for outcome prediction. *BMC Genomics*, 2008, 9 : 117. doi:10.1186/1471-2164-9-177.

90. Kong S.W., Pu W.T. & Park P.J. A multivariate approach for integreting genome-wide expression data and biological knowledge. *Bioinformatics*, 2006, 22(19) : 2373-2380. doi:10.1093/bioinformatics/btl401.

91. Lambert C., Léonard N., De Bolle X. & Depiereux E. ESyPred3D: prediction of proteins 3D structure. *Bioinformatics*, 2002, 18(9) : 1250-1256.

92. Larsson O., Wahlestedt C. & Timmons J.A. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics*, 2005, 6 : 129. doi: 10.1186/1471-2105-6-129.

93. Lewin A. & Grieve I.C. Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. *BMC Bioinformatics*, 2006, 7 : 426. doi:10.1186/1471-2105-7-426.

94. Lewin B. Genes VI. *Oxford University Press & Cell Press*. 1997.

95. Li C. & Wong W.H. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol.*, 2001, 2(8) : research0032.1–research0032.11, *http://genomebiology.com/2001/2/8/research/0032.*

96. Li C. & Wong W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *P. Natl. Acad. Sci. U.S.A.*, 2001, 98 : 31-36

97. Lindley D. V. & Smith A. F. M. Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, B, 1972, 34 : 1–41.

98. Liu Q., Dinu I., Adewale A.J., Potter J.D. & Yasui Y. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 2007, 8 : 431. doi:10.1186/1471-2105-8-431.

99. Long A.D., Mangalam H.J., Chan B.Y., Tolleri L., Hatfield G.W. & Baldi P. Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J. Biol. Chem.*, 2001, 276(23) :19937-44. Epub 2001 Mar 20.

100. Lönnstedt I. & Speed T. Replicated microarray data. *Statistica Sinica*, 2002, 12 : 31–46.

101. Man M.Z., Wang X. & Wang Y. POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, 2000, 16(11) : 953-9.

102. Manda S.O., Walls R.E. & Gilthorpe M.S. A full Bayesian hierarchical mixture model for the variance of gene differential expression. *BMC Bioinformatics*, 2007, 8 : 124.

103. Mann H. B. & Whitney D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, 1947, 18 : 50-60.

104. Mansmann U. & Meister R. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf Med.*, 2005, 44(3) : 449-53.

105. McCall M.N. & Irizarry R.A. Consolidated strategy for the analysis of microarray spike-in data. *Nucleic Acids Res.*, 2008, 36(17) : e108. doi:10.1093/nar/gkn430

106. Mootha V.K., Lindgren C.M., Eriksson K.F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstråle M., Laurila E., Houstis N., Daly M.J., Patterson N., Mesirov J.P., Golub T.R., Tamayo P., Spiegelman B., Lander E.S., Hirschhorn J.N., Altshuler D. & Groop L.C. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 2003, 34(3) : 267-73.

107. Newton M.A., Quintana F.A., Den Boon J.A., Sengupta S. & Ahlquist P. Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *The Annals of Applied Statistics*, 2007, 1(1) : 85-106. doi:10.1214/07-AOAS104.

108. Nilsson R., Peña J.M., Björkegren J. & Tegnér J. Detecting multivariate differentially expressed genes. *BMC Bioinformatics*, 2007, 8 : 150.

109. Opgen-Rhein R. & Strimmer K. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, 2007, 6 : 9.

110. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 2002, 18 : 546-554.

111. Park P.J., Butte A.J. & Kohane I.S. Comparing expression profiles of genes with similar promoter regions. *Bioinformatics*, 2002, 18 : 1576-1584.

112. Parkinson H., Kapushesky M., Shojatalab M., Abeygunawardena N., Coulson R., Farne A., Holloway E., Kolesnykov N., Lilja P., Lukk M., Mani R., Rayner T., Sharma A., William E., Sarkans U. & Brazma A. ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, 2007, 35(Database issue) : D747-750.

113. Pavlidis P., Furey T.S., Liberto M., Haussler D. & Grundy W.N. Promoter region-based classification of genes. *Pac. Symp. Biocomput.*, 2001, 6 : 151-164.

114. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res.*, 2004, 29(6) : 1213-22.

115. Pearson R.D. A comprehensive re-analysis of the Golden Spike data: Towards a benchmark for differential expression methods. *BMC Bioinformatics*, 2008, 9 : 164. doi:10.1186/1471-2105-9-164.

116. Pepper S.D., Saunders E.K., Edwards L.E., Wilson C.L. & Miller C.J. The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics*, 2007, 8 : 273.

117. Ploner A., Miller L.D., Hall P., Bergh J. & Pawitan Y. Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC Bioinformatics*, 2005, 6 : 80.

118. Purdom E. & Holmes S.P. Error Distribution for Gene Expression Data. *Statistical Applications in Genetics and Molecular Biology*, 2005, 4(1):16.

119. Quertemont E. Méthodes Quantitatives en Sciences Psychologiques. *Syllabus des modules de cours STAT0032 et STAT0033*, 2004-05, ULg.

120. Sartor M.A., Tomlinson C.R., Wesselkamper S.C., Sivaganesan S., Leikauf G.D. & Mdevedovic M. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 2006, 7 : 538. doi:10.1186/1471-2105-7-538.

121. Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 1946, 2(6) : 110–114.

122. Saxena V., Orgill D. & Kohane I. Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res.*, 2006, 34 : 22e151. Doi: 10.1093/nar/gkl766.

123. Shaik J.S. & Yeasin M. A unified framework for finding differentially expressed genes from microarray experiments. *BMC Bioinformatics*, 2007, 8 : 347.

124. Smyth G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 2004, 3 : 3.

125. Song S. & Black M.A. Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, 2008, 9 : 502. doi:10.1186/1471-2105-9-502.

126. Spruill S.E., Lu J., Hardy S., Weir B. Assessing sources of variability in microarray gene expression data. *Biotechniques*. 2002, 33(4) : 916-20, 922-3.

127. Stokes A.V. Open standards in medical informatics. *Medinfo.*, 1995, 8 Pt 1 : 177-81.

128. Storey, J.D. A direct approach to false discovery rates. *J. Roy. Statist. Soc. B*, 2002, 64 : 479–498.

129. Storey J.D. & Tibshirani R. Statistical significance for genomewide studies. *PNAS*, 2003, 100(16) : 9440-9445.

130. Student. The Probable Error of a Mean. *Biometrika*, 1908, 1-25.

131. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S. & Mesirov J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 2005,102(43) : 15545-50. Epub 2005 Sep 30.

132. Tian L., Greenberg S.A., Kong S.W., Altschuler J., Kohane I.S. & Park P.J. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U. S. A.*, 2005, 102(38) : 13544-9. Epub 2005 Sep 8.

133. Thomas R.S., Rank D.R., Penn S.G., Zastrow G.M., Hayes K.R., Pande K., Glover E., Silander T., Craven M.W., Reddy J.K., Jovanovich S.B. & Bradfield C.A. Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol Pharmacol.*, 2001, 60(6) : 1189-94.

134. Thomas J.G., Olson J.M., Tapscott S.J. & Zhao L.P. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, 2001, 11(7) : 1227-36.

135. Trajkovski I., Lavrac N. & Tolar J. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 2008, 41: 588-601, doi:10.1016/j.jbi.2007.12.001.

136. Tusher V.G., Tibshirani R. & Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *P. Natl. Acad. Sci. U.S.A.*, 2001, 98 : 5116-5121.

Erratum in: *P. Natl. Acad. Sci. U.S.A.*, 2001, 98 : 10515.

137. van de Wiel MA, Smeets SJ, Brakenhoff RH, Ylstra B. CGHMultiArray: exact P-values for multi-array comparative genomic hybridization data. *Bioinformatics*, 2005, 21(14) : 3193-4. Epub 2005 May 6.

138. van Helden J., Rios A.F. & Collado-Vides J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, 2000, 281 : 1808-1818.

139. Vengellur A., Phillips J.M., Hogenesch J.B. & LaPres J.J. Gene expression profiling of hypoxia signaling in human hepatocellular carcinoma cells. *Physiol. Genomics*, 2005, 22 : 308-318.

140. Virtaneva K., Wright F.A., Tanner S.M., Yuan B., Lemon W.J., Caligiuri M.A., Bloomfield C.D., de La Chapelle A. & Krahe R. Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl. Acad. Sci. U. S. A.*, 2001, 98(3) : 1124-9.

141. Watson J.D., Gilman M., Witkowski J. & Zoller M. Recombinant DNA. *Freeman*, 1992.

142. Watson M.D. CoXpress: differential co-expression in gene expression data. *BMC Bioinformatics*, 2006, 7 : 509. doi:10.1186/1471-2105-7-509.

143. Welch B.L. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 1938, 29 : 350-362.

144. Westfall, P. & Young, S. Resampling-Based Multiple Testing. *Wiley*, 1993, New York

145. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics*, 1945, 1 : 80-83.

146. Wu Z., Irizarry R.A., Gentleman R., Murillo F.M. & Spencer F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.*, 2004, 99 : 909-917.

147. Wu L., Williams P.M. & Koch W. Clinical applications of microarray-based diagnostic tests. *Biotechniques*. 2005, 39(10 Suppl) : S577-82.

148. Xiao Y., Segal M.R., Rabert D., Ahn A.H., Anand P., Sangameswaran L., Hu D. & Hunt C.A. Assessment of differential gene expression in human peripheral nerve injury. *BMC Genomics*. 2002, 3(1) : 28.

149. Yap Y.L., Lam D.C., Luc G., Zhang X.W., Hernandez D., Gras R., Wang E., Chiu S.W., Chung L.P., Lam W.K., Smith D.K., Minna J.D., Danchin A. & Wong M.P. Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucl. Ac. Res.*, 2005, 33 : 409-421.

Erratum in: *Nucl. Ac. Res.*, 2005, 33 : 2764.

150. Zahn J.M., Sonu R., Vogel H., Crane E., Mazan-Mamezarz K., Rabkin R., Davis R.W., Becker K.G., Owen A.B. & Kim S.K. Transcriptional profiling of aging in human muscle reveals a common aging structure. *PloS Genetics*, 2006, 2 : 1058-1069.

151. Zhang B., Schmoyer D., Kirov S. & Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 2004, 5 : 16.

152. Zhang S. A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC Bioinformatics*, 2007, 8 : 230. doi:10.1186/1471-2105-8-230.

VIII. Annexes

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-value	Consensus (rank)	Weighted Consensus (rank)	Probeset
ABCF2	-	-	-	-	-	-	55	-	-	-	-	-	-	207622_s_at
ACLY	-	-	-	-	-	-	41	-	-	-	-	-	-	210337_s_at
ACP2	-	96	-	70	-	-	-	-	-	-	-	95	95	202767_at
ADAM10	-	-	-	-	-	-	45	40	-	-	-	-	-	202604_x_at
ADAM8	-	-	-	96	-	-	-	-	-	-	-	-	-	205180_s_at
ADCY7	-	-	-	-	-	-	93	-	-	-	-	-	-	203741_s_at
ADFP	96	14	82	5	1	91	72	-	23	14	10	8	4	209122_at
ADM	**12**	**-**	**9**	**97**	**73**	**12**	**-**	**43**	**-**	**41**	**58**	**28**	**43**	**202912_at**
AK3L1	**-**	**52**	**-**	**48**	**68**	**-**	**-**	**-**	**-**	**-**	**98**	**-**	**81**	**204348_s_at**
AKR1C1	-	-	-	-	-	-	-	-	52	-	-	-	-	204151_x_at
AKT3	-	-	-	-	-	-	-	85	-	-	-	-	-	212609_s_at
ALCAM	-	-	-	-	-	-	-	84	-	-	-	-	-	201951_at
ALDH9A1	-	-	-	-	-	-	75	-	-	-	-	-	-	201612_at
ALDOA	**-**	**90**	**-**	**80**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**214687_x_at**
ALDOC	**-**	**54**	**-**	**42**	**34**	**-**	**-**	**-**	**100**	**66**	**57**	**61**	**56**	**202022_at**
AP2A2	-	-	-	-	-	-	-	92	-	-	-	-	-	212159_x_at
APOBEC3A	-	-	-	-	-	-	-	-	91	-	-	-	-	210873_x_at
ARHGAP25	94	-	59	-	-	94	-	-	-	-	-	-	-	204882_at
ARNT2	-	-	-	-	-	-	-	-	81	-	-	-	-	202986_at
ASCC1	70	-	33	-	-	68	-	-	-	-	-	-	-	219336_s_at
ATP2C1	29	-	25	-	-	33	-	-	-	-	-	-	-	212255_s_at
ATP7A	-	-	-	-	-	-	54	-	-	-	-	-	-	205198_s_at
BCAT1	**-**	**81**	**-**	**-**	**88**	**-**	**-**	**-**	**-**	**-**	**-**	**99**	**86**	**214390_s_at**
	-	**-**	**-**	**-**	**75**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**214452_at**
BHLHB2	**-**	**-**	**-**	**-**	**54**	**-**	**-**	**46**	**-**	**64**	**87**	**66**	**100**	**201169_s_at**
	-	**-**	**-**	**-**	**98**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**201170_s_at**
BNIP3	**31**	**6**	**66**	**7**	**24**	**28**	**-**	**-**	**4**	**5**	**4**	**6**	**6**	**201848_s_at**
	91	**10**	**-**	**22**	**43**	**84**	**-**	**-**	**5**	**9**	**6**	**30**	**23**	**201849_at**
BNIP3L	**78**	**47**	**-**	**30**	**36**	**72**	**-**	**-**	**-**	**62**	**54**	**54**	**47**	**221478_at**
	63	**40**	**87**	**28**	**27**	**56**	**-**	**-**	**-**	**80**	**61**	**58**	**46**	**221479_s_at**
BRD4	-	-	-	-	-	-	-	47	-	-	-	-	-	202103_at
C3orf28	46	91	26	-	20	44	-	-	58	51	49	43	51	220942_x_at
C8orf70	-	-	-	-	-	-	64	24	-	-	-	-	-	205308_at
C9orf114	-	87	-	76	-	-	-	-	-	-	-	-	-	218565_at
CA12	**-**	**83**	**-**	**53**	**-**	**-**	**-**	**-**	**74**	**-**	**-**	**-**	**-**	**203963_at**
	-	**-**	**-**	**67**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**210735_s_at**
CAMKK2	-	-	-	-	-	-	-	90	-	-	-	-	-	212252_at
CAPN7	-	-	-	-	-	-	-	74	-	-	-	-	-	203357_s_at
CCL18	-	-	-	-	-	-	-	-	80	-	-	-	-	209924_at
	-	-	-	-	-	-	25	11	37	59	77	-	-	32128_at
CCL20	**-**	**42**	**-**	**46**	**69**	**-**	**-**	**-**	**14**	**34**	**30**	**65**	**59**	**205476_at**
CCL23	**98**	**43**	**-**	**83**	**17**	**96**	**-**	**-**	**49**	**49**	**44**	**51**	**44**	**210548_at**
CD163	-	57	-	51	95	-	-	-	-	-	97	-	82	203645_s_at
	-	68	-	52	57	-	42	-	-	54	59	56	60	215049_x_at
	-	71	-	-	-	-	-	-	30	98	79	-	-	216233_at
CD300A	**-**	**28**	**-**	**35**	**28**	**100**	**-**	**-**	**69**	**52**	**46**	**48**	**37**	**209933_s_at**
	-	**60**	**-**	**-**	**-**	**-**	**-**	**-**	**53**	**87**	**84**	**-**	**-**	**217078_s_at**
CD36	-	-	-	-	-	-	-	-	85	-	-	-	-	206488_s_at
	-	98	-	-	-	-	-	-	31	-	93	-	-	209555_s_at
CD84	-	-	-	-	-	-	-	-	42	-	-	-	-	205988_at
	-	-	-	-	-	-	-	-	61	-	-	-	-	211192_s_at
CD86	-	-	-	-	-	-	88	-	-	100	-	93	-	210895_s_at
CDC42EP3	-	-	-	-	-	-	65	-	-	-	-	-	-	209287_s_at
CFLAR	-	-	-	-	-	-	31	60	-	-	-	-	-	211862_x_at
CHI3L1	-	89	-	61	61	-	-	-	66	-	86	-	93	209395_at
	-	-	-	-	76	-	38	-	43	56	68	82	-	209396_s_at
CLCN7	-	-	-	-	-	-	14	59	-	-	-	-	-	221961_at
CMTM6	-	-	53	-	-	-	-	-	-	-	-	-	-	217947_at
CNIH	-	-	-	-	-	-	-	91	-	-	-	-	-	201653_at
CNOT8	-	-	-	-	-	-	95	31	-	-	-	-	-	202164_s_at

Annexe I - page 1: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-val	Consensus (rank)	Weighted Consensus (rank)	Probeset
COPS2	48	-	36	-	-	51	-	-	-	-	-	-	-	202467_s_at
CRKL	-	84	-	-	-	-	-	-	-	-	-	-	-	206184_at
CRLF3	-	-	-	-	-	-	90	83	-	-	-	-	-	205474_at
CSF3	-	-	-	-	-	-	-	-	98	-	-	-	-	207442_at
CSNK1A1	-	-	-	-	-	-	16	95	-	-	-	-	-	213860_x_at
CTNNA1	20	-	11	-	-	36	-	-	-	-	-	-	-	210844_x_at
CTSC	**-**	**-**	**-**	**-**	**80**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**201487_at**
CTSO	-	-	-	-	-	-	49	-	-	-	-	-	-	203758_at
CXCL5	-	-	-	87	-	-	-	-	-	-	-	-	-	215101_s_at
CXCR7	-	-	-	-	-	-	-	-	21	-	91	-	-	212977_at
DAPK1	-	100	-	60	-	-	-	-	-	-	-	-	-	203139_at
DAPK3	-	-	-	-	-	-	97	30	-	-	-	-	-	203891_s_at
DARS	-	-	-	-	84	-	-	-	-	93	-	92	-	201623_s_at
	-	-	-	91	93	-	-	-	-	-	-	-	-	201624_at
DDIT4	**-**	**-**	**-**	**-**	**91**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**202887_s_at**
DEGS1	-	-	-	-	-	-	99	-	-	-	-	-	-	207431_s_at
DENND3	-	-	99	-	-	-	-	-	-	-	-	97	-	212974_at
DFNA5	-	*67*	-	*66*	*79*	-	*70*	-	-	*79*	*80*	*74*	*71*	*203695_s_at*
DHCR24	-	-	-	88	-	-	-	-	-	-	-	-	-	200862_at
DRAM	-	-	-	-	-	-	39	28	-	-	-	-	-	218627_at
EGLN1	**27**	**19**	**22**	**9**	**42**	**26**	**-**	**-**	**89**	**38**	**37**	**25**	**22**	**221497_x_at**
EGLN3	-	-	-	-	-	-	-	-	*84*	-	-	-	-	*219232_s_at*
EGR1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**50**	**-**	**-**	**-**	**-**	**201694_s_at**
ELK1	83	-	38	-	-	78	-	-	-	-	-	-	-	203617_x_at
EMR1	-	-	-	-	-	-	83	36	-	-	-	-	-	207111_at
ENO1	**4**	**35**	**3**	**21**	**31**	**4**	**-**	**-**	**-**	**23**	**36**	**10**	**15**	**201231_s_at**
	52	**38**	**49**	**38**	**40**	**46**	**-**	**-**	**-**	**55**	**53**	**41**	**38**	**217294_s_at**
ENO2	**76**	**8**	**-**	**41**	**48**	**67**	**-**	**-**	**7**	**11**	**7**	**34**	**25**	**201313_at**
ENOSF1	-	-	-	-	-	-	-	-	93	-	-	-	-	204143_s_at
	-	-	-	-	-	-	-	-	28	-	-	-	-	213645_at
ENPP2	-	58	-	36	37	-	-	-	33	47	40	60	52	209392_at
	-	36	-	24	33	-	37	9	20	13	13	24	28	210839_s_at
ERO1L	**24**	**48**	**19**	**39**	**30**	**21**	**-**	**-**	**-**	**46**	**48**	**31**	**32**	**218498_s_at**
EVL	-	-	-	-	-	-	76	-	-	-	-	-	-	217838_s_at
FAM21C	-	-	-	-	-	-	85	-	-	-	-	-	-	212929_s_at
FBXO21	-	-	-	-	-	-	47	19	-	-	-	-	-	212231_at
FCAR	95	-	65	-	-	95	-	-	-	-	-	-	-	211816_x_at
FCGR1A	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**40**	**-**	**-**	**-**	**-**	**216950_s_at**
FCGR1B	-	77	-	84	63	-	-	-	12	36	32	79	69	214511_x_at
FCGR2B	**-**	**51**	**-**	**34**	**39**	**-**	**-**	**-**	**-**	**67**	**64**	**59**	**57**	**210889_s_at**
FCGR2C	45	-	77	74	86	43	-	-	-	-	-	73	75	210992_x_at
	22	-	44	-	59	19	-	-	-	60	78	49	65	211395_x_at
FLJ10815	-	-	-	-	-	-	35	-	-	-	-	-	-	218727_at
FLJ20323	-	-	-	-	-	-	92	-	-	-	-	-	-	211724_x_at
FLNB	-	-	-	-	-	-	-	75	-	-	-	-	-	208613_s_at
FLT1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**96**	**-**	**100**	**-**	**210287_s_at**
	-	**72**	**-**	**-**	**-**	**-**	**50**	**17**	**36**	**48**	**55**	**83**	**-**	**222033_s_at**
FN1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**70**	**-**	**-**	**-**	**-**	**210495_x_at**
	-	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**99**	**-**	**-**	**-**	**-**	**211719_x_at**
FUCA1	-	-	-	-	-	-	-	-	82	-	-	-	-	202838_at
G3BP1	62	-	37	-	-	64	-	-	-	-	-	-	-	201503_at
G6PD	-	-	-	-	-	-	-	48	-	-	-	-	-	202275_at
GAPDH	**2**	**-**	**12**	**-**	**85**	**2**	**63**	**67**	**-**	**43**	**63**	**23**	**40**	**212581_x_at**
	23	**69**	**13**	**81**	**56**	**22**	**-**	**-**	**-**	**72**	**75**	**50**	**54**	**213453_x_at**
	11	**-**	**5**	**99**	**-**	**13**	**-**	**-**	**-**	**69**	**85**	**47**	**62**	**217398_x_at**
	80	**-**	**51**	**-**	**-**	**83**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**AFFX-HUMGAPDH/M33197_3_at**
	3	**-**	**4**	**-**	**70**	**3**	**-**	**-**	**-**	**50**	**69**	**27**	**42**	**AFFX-HUMGAPDH/M33197_5_at**
	9	**-**	**23**	**-**	**83**	**8**	**11**	**73**	**-**	**42**	**66**	**32**	**53**	**AFFX-HUMGAPDH/M33197_M_at**

Annexe I - page 2: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-val	Consensus (rank)	Weighted Consensus (rank)	Probeset
GAS7	-	*46*	-	-	-	-	-	-	*56*	*73*	*71*	*91*	*84*	*202191_s_at*
	-	*44*	-	-	-	-	-	-	*46*	*76*	*67*	-	*97*	*202192_s_at*
	-	-	-	-	-	-	*22*	*77*	*29*	*53*	*65*	-	-	*207704_s_at*
	-	*64*	-	-	-	-	*71*	-	*6*	*19*	*17*	*89*	*83*	*210872_x_at*
	-	*94*	-	-	-	-	-	-	*35*	*77*	*73*	-	-	*211067_s_at*
GBA	81	-	54	-	-	85	-	-	-	-	-	-	-	210589_s_at
GBE1	**77**	**-**	**-**	**-**	**64**	**70**	**-**	**-**	**-**	**-**	**-**	**86**	**-**	**203282_at**
GIMAP5	67	-	-	-	-	63	-	-	-	-	-	-	-	218805_at
GLUD1	-	-	-	-	-	-	53	15	-	-	-	-	-	200946_x_at
GPI	**93**	**41**	**67**	**29**	**9**	**88**	**-**	**-**	**34**	**30**	**29**	**33**	**27**	**208308_s_at**
GPNMB	-	-	-	-	99	-	-	-	94	-	-	-	-	201141_at
GPR107	100	-	72	-	-	-	-	-	-	-	-	-	-	211979_at
GPR109B	38	45	21	71	41	37	-	-	-	57	56	40	45	205220_at
GPX3	-	-	-	-	-	-	-	96	-	-	-	-	-	201348_at
GRLF1	-	-	-	-	-	-	9	81	-	-	-	-	-	202046_s_at
GYS1	-	-	-	-	58	-	-	-	-	-	-	-	94	201673_s_at
HEATR1	-	-	-	-	-	-	-	61	-	-	-	-	-	218594_at
HIG2	**6**	**1**	**20**	**2**	**7**	**5**	**-**	**-**	**3**	**3**	**3**	**1**	**1**	**218507_at**
HIPK2	-	-	-	-	-	-	-	57	-	-	-	-	-	213763_at
HP1BP3	-	-	-	-	-	-	-	49	-	-	-	-	-	220633_s_at
HS3ST1	-	29	74	13	100	-	-	-	-	71	60	55	49	205466_s_at
HSPA6	-	-	-	-	*77*	-	*78*	-	-	-	-	-	-	*213418_at*
IBRDC3	-	-	-	-	-	-	89	99	-	-	-	-	-	213038_at
IFI44	-	-	-	-	-	-	-	62	-	-	-	-	-	214453_s_at
IFT122	-	99	-	-	-	-	-	-	-	-	-	-	-	220744_s_at
IGBP1	61	-	32	-	-	66	-	-	-	-	-	-	-	202105_at
IGHG1	-	-	-	-	-	-	-	-	*47*	-	-	-	-	*213674_x_at*
	68	-	*91*	-	-	*75*	-	-	-	-	-	-	-	*217039_x_at*
IGSF6	-	-	-	-	-	-	80	-	-	-	-	-	-	206420_at
IL1A	**-**	**-**	**-**	**90**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**210118_s_at**
IL1RN	**-**	**82**	**92**	**79**	**66**	**-**	**-**	**-**	**-**	**94**	**96**	**75**	**73**	**212659_s_at**
	89	**-**	**57**	**-**	**53**	**86**	**-**	**-**	**-**	**91**	**95**	**71**	**76**	**216243_s_at**
IMPA2	-	39	60	75	81	-	-	-	-	-	-	-	80	203126_at
INHBA	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**57**	**-**	**-**	**-**	**-**	**210511_s_at**
INSIG1	-	-	-	-	-	-	-	58	-	-	-	-	-	201627_s_at
JMJD1A	**-**	**63**	**-**	**65**	**49**	**-**	**-**	**-**	**-**	**-**	**88**	**84**	**72**	**212689_s_at**
KDELR2	72	-	62	-	-	90	-	-	-	-	-	-	-	200698_at
KIAA0251	-	-	-	-	-	-	-	63	-	-	-	-	-	212053_at
KLF10	-	-	-	-	82	-	-	-	71	-	-	-	-	202393_s_at
KLHL18	*57*	*15*	-	*6*	-	*60*	-	-	-	-	*81*	*70*	*55*	*212882_at*
LAD1	-	-	-	-	-	-	-	100	-	-	-	-	-	203287_at
LAMP3	-	-	-	-	-	-	-	78	-	-	-	-	-	205569_at
LASP1	*21*	-	*24*	-	-	*25*	-	-	-	-	-	-	-	*200618_at*
LGALS2	-	-	-	-	-	-	-	-	90	-	-	-	-	208450_at
LGALS8	*18*	-	*68*	*63*	*72*	*17*	*5*	*12*	-	*21*	*42*	*19*	*33*	*208934_s_at*
	60	-	*29*	-	*96*	*54*	-	-	-	*90*	-	*67*	*78*	*208936_x_at*
	19	-	*35*	-	*90*	*18*	-	-	-	*85*	-	*63*	*79*	*210732_s_at*
LHFPL2	-	-	85	-	-	-	-	-	-	-	-	-	-	212658_at
LIPA	-	-	-	-	-	-	-	-	55	-	-	-	-	201847_at
LITAF	-	-	-	-	-	-	-	53	-	-	-	-	-	200706_s_at
LOH11CR2A	-	-	-	-	-	-	15	86	-	-	-	-	-	210102_at
LONP1	-	-	93	-	71	-	-	-	-	-	-	94	90	209017_s_at
LXN	-	-	-	-	-	-	96	-	-	-	-	-	-	218729_at
MAFB	-	-	70	-	-	-	-	-	-	-	-	-	-	218559_s_at
MAN1A2	-	-	-	-	-	-	-	68	-	-	-	-	-	217922_at
MAPK13	-	-	-	-	-	-	-	51	-	-	-	-	-	210058_at
MAPK7	59	-	-	-	-	59	-	-	-	-	-	-	-	35617_at
MATR3	8	-	8	-	-	10	17	-	-	83	-	85	-	200624_s_at
MERTK	*39*	*4*	*-*	*3*	*8*	*35*	*6*	*23*	*15*	*4*	*5*	*2*	*2*	*206028_s_at*
	44	*5*	*41*	*4*	*5*	*42*	-	-	*19*	*12*	*8*	*7*	*3*	*211913_s_at*

Annexe I - page 3: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-va	Consensus (rank)	Weighted Consensus (rank	Probeset
METAP1	*16*	-	*6*	-	-	*20*	-	-	-	-	-	-	-	*212673_at*
MFSD1	-	-	79	-	-	-	-	-	-	-	-	-	-	218109_s_at
MGC5139	-	-	-	-	-	-	26	-	-	-	-	-	-	202365_at
MICA	-	-	78	-	-	-	-	-	-	-	-	-	-	205905_s_at
MIF	**-**	**18**	**-**	**11**	**22**	**-**	**-**	**-**	**24**	**35**	**19**	**42**	**30**	**217871_s_at**
MKKS	-	-	-	-	-	-	66	-	-	-	-	-	-	218138_at
MLX	79	-	96	-	-	80	-	-	-	-	-	-	-	217909_s_at
MMP1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**18**	**78**	**70**	**-**	**-**	**204475_at**
MMP19	-	-	-	-	-	-	27	-	-	-	-	-	-	204574_s_at
MNDA	-	74	-	-	-	-	-	-	73	-	-	-	-	204959_at
MOAP1	40	-	-	-	-	62	-	-	-	-	-	-	-	212508_at
MRPS10	-	-	-	-	-	-	19	69	-	-	-	-	-	218106_s_at
MTF2	87	-	40	-	-	97	-	-	-	-	-	-	-	209705_at
MX2	-	-	-	-	-	-	-	80	-	-	-	-	-	204994_at
MXI1	**-**	**31**	**-**	**15**	**29**	**-**	**-**	**-**	**32**	**40**	**34**	**45**	**35**	**202364_at**
MYH11	-	-	-	-	-	-	-	-	41	-	-	-	-	201497_x_at
MYO1B	-	-	-	98	-	-	-	-	-	-	-	-	-	212365_at
MZF1	50	-	34	-	-	53	-	-	-	-	-	-	-	204139_x_at
NDRG1	**-**	**73**	**84**	**57**	**32**	**-**	**-**	**-**	**-**	**84**	**76**	**64**	**66**	**200632_s_at**
NEDD4L	-	86	-	-	-	-	-	-	-	-	-	-	-	212445_s_at
NEU1	-	-	-	-	-	-	-	41	-	-	-	-	-	208926_at
NID1	*86*	-	*63*	-	-	*81*	-	-	-	-	-	-	-	*202007_at*
	-	-	-	-	-	-	*10*	*18*	-	-	-	-	-	*202008_s_at*
NOTCH3	-	-	-	-	-	-	-	50	-	-	-	-	-	203238_s_at
NPEPPS	-	-	-	-	-	-	20	-	-	-	-	-	-	201455_s_at
NR1H3	-	*49*	-	*68*	-	-	*3*	*3*	*9*	*6*	*11*	*35*	*48*	*203920_at*
NR4A2	-	-	-	-	-	-	-	-	83	-	-	-	-	204621_s_at
NT5E	-	-	-	-	-	-	62	34	11	33	39	-	-	203939_at
NUPL1	-	-	-	-	-	-	-	71	-	-	-	-	-	204435_at
NUTD3/RPS10	-	-	-	-	-	-	91	-	-	-	-	-	-	211976_at
OAS1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**97**	**-**	**-**	**-**	**-**	**202869_at**
OLFML2B	-	*11*	-	*23*	*35*	*99*	-	-	*26*	*20*	*14*	*26*	*24*	*213125_at*
OXSM	-	-	-	-	-	-	46	13	-	-	-	-	-	219133_at
P4HA1	**51**	**55**	**27**	**33**	**18**	**45**	**12**	**5**	**95**	**16**	**24**	**9**	**18**	**207543_s_at**
P4HA2	**-**	**70**	**-**	**-**	**-**	**-**	**-**	**-**	**60**	**-**	**-**	**-**	**-**	**202733_at**
PAIP1	-	-	-	-	-	-	-	65	-	-	-	-	-	209064_x_at
PAM	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**96**	**-**	**-**	**-**	**-**	**202336_s_at**
PANX1	-	-	-	*77*	-	-	*84*	*38*	-	*82*	*89*	*88*	*96*	*204715_at*
PAPOLA	-	-	98	-	-	-	-	-	-	-	-	-	-	212720_at
PARVB	-	*59*	-	*37*	*89*	-	-	-	*45*	*99*	*72*	-	*88*	*37966_at*
PCCB	-	-	-	-	-	-	-	88	-	-	-	-	-	212694_s_at
PCDHGC3	-	92	-	94	-	-	-	-	-	-	-	-	-	205717_x_at
PDK1	**7**	**30**	**39**	**-**	**52**	**6**	**-**	**-**	**25**	**17**	**16**	**14**	**21**	**206686_at**
PDK3	-	53	-	-	-	-	-	-	-	-	-	-	-	221957_at
PDLIM5	-	-	-	-	-	-	-	98	-	-	-	-	-	212412_at
PFKP	**88**	**12**	**-**	**47**	**12**	**82**	**-**	**-**	**64**	**61**	**43**	**52**	**34**	**201037_at**
PFTK1	-	*32*	-	*45*	-	-	-	-	-	-	*94*	*78*	*70*	*211502_s_at*
PGAM1	**34**	**-**	**17**	**-**	**-**	**34**	**-**	**-**	**-**	**-**	**-**	**87**	**-**	**200886_s_at**
PGK1	**33**	**-**	**61**	**93**	**45**	**30**	**-**	**-**	**-**	**75**	**74**	**57**	**61**	**200737_at**
	1	**-**	**1**	**64**	**51**	**1**	**7**	**2**	**-**	**7**	**28**	**3**	**13**	**200738_s_at**
	26	**-**	**10**	**-**	**50**	**24**	**-**	**-**	**-**	**68**	**82**	**53**	**67**	**217356_s_at**
PGM1	*15*	*9*	*71*	*17*	*38*	*14*	-	-	*13*	*10*	*9*	*12*	*10*	*201968_s_at*
PHC2	99	-	47	-	-	-	-	-	-	-	-	-	-	200919_at
PHKB	-	-	-	-	-	-	29	64	-	-	-	-	-	202738_s_at
PHLDA1	-	-	-	-	-	-	68	27	-	-	-	-	-	217999_s_at
PIH1D1	5	-	2	-	-	7	48	-	-	88	-	-	-	217872_at
PIK3CB	-	97	-	-	-	-	-	-	-	-	-	-	-	217620_s_at
PLA2G4B	-	-	-	-	-	-	30	7	-	-	-	-	-	219095_at
PLAUR	-	-	-	-	-	-	59	76	-	-	-	-	-	211924_s_at
PLEKHG3	-	62	-	82	-	-	-	-	-	-	-	-	-	212823_s_at

Annexe I - page 4: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-val	Consensus (rank)	Weighted Consensus (rank)	Probeset
PLOD2	-	-	-	-	-	-	-	-	17	-	90	-	-	202619_s_at
	-	-	-	-	-	-	-	-	8	65	50	-	-	202620_s_at
PLXND1	-	-	-	-	-	-	58	-	-	-	-	-	-	38671_at
PPIF	*92*	-	*42*	-	-	*98*	-	-	-	-	-	-	-	*201489_at*
PPP3CA	-	-	-	-	-	-	74	-	-	-	-	-	-	202457_s_at
PPP4C	-	-	-	-	-	-	82	87	-	-	-	-	-	208932_at
PRDX3	41	-	48	-	-	50	-	-	-	-	-	-	-	201619_at
PRDX4	-	*85*	-	*59*	*87*	-	-	-	-	-	*100*	*81*	*77*	*201923_at*
PRKD3	-	-	-	-	-	-	44	33	-	-	-	-	-	211084_x_at
PRO1843	-	-	-	-	-	-	79	29	-	-	-	-	-	219599_at
PRSS1	-	-	-	-	-	-	33	14	-	-	-	-	-	211796_s_at
PTGS2	**-**	**-**	**-**	**78**	**-**	**-**	**-**	**-**	**68**	**-**	**-**	**-**	**-**	**204748_at**
QKI	-	-	-	-	-	-	-	89	-	-	-	-	-	212262_at
RAB20	-	-	-	-	-	-	86	-	-	-	-	-	-	219622_at
RAB2A	-	-	-	-	-	-	51	-	-	-	-	-	-	208732_at
RAB5C	-	-	-	-	-	-	52	16	-	-	-	-	-	201156_s_at
RANBP5	49	-	-	-	-	49	-	-	-	-	-	-	-	211955_at
RANBP9	-	-	-	-	-	-	73	56	-	-	-	-	-	202583_s_at
RBM4B	-	-	-	-	-	-	61	22	-	-	-	-	-	209497_s_at
RCBTB1	-	-	55	-	-	-	-	-	-	-	-	-	-	218352_at
RGC32	-	-	-	-	-	-	-	-	88	-	-	-	-	218723_s_at
RMI1	56	-	58	-	-	55	-	66	-	-	-	98	-	218979_at
RNASET2	*35*	*56*	*94*	*54*	*13*	*32*	-	-	-	*58*	*51*	*39*	*36*	*217983_s_at*
	36	*22*	-	*32*	*14*	*31*	-	-	*59*	*45*	*38*	*36*	*29*	*217984_at*
ROCK1	-	-	-	-	-	-	94	-	-	-	-	-	-	216621_at
RPS2	-	-	-	-	-	-	4	1	-	-	-	-	-	221798_x_at
RRM1	-	-	-	-	-	-	23	6	-	-	-	-	-	201476_s_at
RSBN1	-	-	-	-	-	-	21	-	-	-	-	-	-	213694_at
RTCD1	-	-	-	-	-	-	36	42	-	-	-	-	-	203594_at
RTF1	-	-	-	-	-	-	32	-	-	-	-	-	-	212301_at
RTP4	-	-	-	-	-	-	1	4	-	89	-	-	-	219684_at
RXRA	*55*	-	*28*	-	-	*58*	-	-	-	-	-	-	-	*202449_s_at*
S100A9	-	-	83	-	-	-	-	-	-	-	-	-	-	203535_at
SAE2	-	-	-	-	-	-	*43*	*52*	-	*92*	-	*90*	-	*201177_s_at*
SASH1	37	-	31	-	-	38	-	-	-	-	-	-	-	213236_at
SDCBP	*47*	-	-	-	-	*61*	-	-	-	-	-	-	-	*200958_s_at*
SEMA4D	73	-	-	-	-	69	-	-	-	-	-	-	-	203528_at
SGPL1	-	-	-	-	-	-	8	35	-	-	-	-	-	212321_at
	-	-	-	-	-	-	-	70	-	-	-	-	-	212322_at
SIGLEC1	-	-	-	-	-	-	-	-	87	-	-	-	-	219519_s_at
SLC2A3	**17**	**26**	**7**	**27**	**2**	**16**	**-**	**-**	**77**	**31**	**31**	**15**	**8**	**202497_x_at**
	54	**27**	**45**	**16**	**6**	**47**	**-**	**-**	**48**	**26**	**22**	**21**	**16**	**202498_s_at**
	53	**24**	**80**	**14**	**16**	**48**	**98**	**-**	**65**	**25**	**25**	**22**	**20**	**202499_s_at**
	14	**16**	**18**	**8**	**11**	**15**	**-**	**-**	**62**	**24**	**20**	**16**	**9**	**216236_s_at**
	28	**33**	**86**	**44**	**4**	**27**	**40**	**44**	**44**	**18**	**18**	**11**	**14**	**222088_s_at**
SLC2A5	-	95	-	-	-	-	-	-	-	-	-	-	-	204429_s_at
SLC35D1	-	-	-	86	-	-	-	-	-	-	-	-	-	209712_at
SLC36A1	-	-	76	-	97	-	-	-	-	86	-	77	99	213119_at
SLC7A11	-	-	-	-	-	-	57	25	-	-	-	-	-	209921_at
SLCO2B1	*30*	*2*	*15*	*1*	*65*	*29*	-	-	*92*	*27*	*21*	*13*	*7*	*203473_at*
	90	-	*52*	-	-	*89*	-	-	-	-	-	-	-	*211557_x_at*
SNAPC1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**79**	**-**	**-**	**-**	**-**	**205443_at**
SPP1	**-**	**21**	**-**	**72**	**94**	**-**	**-**	**-**	**1**	**1**	**1**	**46**	**39**	**209875_s_at**
SSR1	-	-	-	-	-	-	-	45	-	-	-	-	-	200889_s_at
ST14	-	-	-	-	-	-	34	10	-	-	-	-	-	202005_at
STAB1	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**76**	**-**	**-**	**-**	**-**	**204150_at**
	-	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**78**	**-**	**-**	**-**	**-**	**38487_at**
STK38	*65*	-	*90*	-	-	*74*	-	-	-	-	-	-	-	*202951_at*
STOM	-	-	-	-	-	-	-	97	-	-	-	-	-	201061_s_at
SUOX	-	-	-	85	-	-	-	-	-	-	-	-	-	204067_at

Annexe I - page 5: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Gene Symbol	Student t-test	Window t-test	Welch t-test	Window Welch t-test	Regularized t-test	SAM test	Robust Student t-test	Robust Welch t-test	LPE test	Consensus (p-value)	Weighted Consensus (p-val	Consensus (rank)	Weighted Consensus (rank)	Probeset
SYNCRIP	82	-	89	-	-	87	-	-	-	-	-	-	-	209024_s_at
SYNJ1	75	-	100	-	-	77	-	-	-	-	-	-	-	212990_at
SYPL1	66	-	-	-	-	65	-	-	-	-	-	-	-	201259_s_at
TBC1D13	-	-	-	-	-	-	28	20	-	-	-	-	-	44696_at
TCEB3	-	-	-	-	-	-	77	54	-	-	-	-	-	202818_s_at
TESC	-	25	-	18	46	-	-	-	-	-	-	-	92	218872_at
TFPI2	-	88	-	55	62	-	-	-	-	-	-	-	-	209278_s_at
TMEM158	-	*66*	-	*56*	-	-	-	-	*10*	*29*	*26*	*80*	*74*	*213338_at*
TMEM43	*71*	-	*50*	-	-	*76*	-	-	-	-	-	-	-	*217795_s_at*
TMEM45A	-	-	-	-	-	-	-	-	22	97	83	-	-	219410_at
TMEM70	-	-	-	-	-	-	56	-	-	-	-	-	-	219449_s_at
TNFSF13	-	-	-	-	-	-	-	93	-	-	-	-	-	209500_x_at
TNIK	-	78	-	50	-	-	-	-	-	-	-	-	-	213109_at
TNS1	*42*	*7*	-	*25*	*67*	*40*	-	-	*16*	*15*	*12*	*20*	*17*	*218864_at*
	64	*76*	*81*	*62*	*26*	*57*	-	-	*72*	*44*	*47*	*38*	*41*	*221246_x_at*
	13	*93*	*14*	-	*15*	*11*	-	*55*	*63*	*22*	*35*	*17*	*26*	*221747_at*
	43	*20*	*30*	*12*	*3*	*41*	-	-	*67*	*28*	*23*	*18*	*12*	*221748_s_at*
TPI1	**-**	**75**	**69**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**96**	**85**	**200822_x_at**
	58	**80**	**-**	**49**	**60**	**52**	**-**	**-**	**-**	**95**	**92**	**68**	**68**	**213011_s_at**
TRAF3IP2	-	-	-	-	-	-	18	8	-	-	-	-	-	215411_s_at
TREM1	-	-	-	89	-	-	-	-	-	-	-	-	-	219434_at
TRIM14	-	-	-	-	-	-	-	94	-	-	-	-	-	203148_s_at
TRIM37	-	-	-	-	-	-	69	-	-	-	-	-	-	213009_s_at
TRIT1	-	-	88	-	-	-	67	37	-	70	99	62	91	218617_at
TXNIP	-	50	-	31	44	-	-	-	86	81	62	72	64	201008_s_at
	-	-	-	-	55	-	-	-	-	-	-	-	-	201009_s_at
	-	65	-	40	92	-	-	-	-	-	-	-	98	201010_s_at
UBE2A	-	-	95	-	-	-	-	-	-	-	-	-	-	201899_s_at
USP10	-	-	97	-	-	-	-	-	-	-	-	-	-	209137_s_at
USP18	-	-	-	-	-	-	60	72	-	-	-	-	-	219211_at
VDAC1	*84*	-	*75*	-	-	*79*	-	-	-	-	-	-	-	*212038_s_at*
VDR	-	-	-	-	-	-	-	79	-	-	-	-	-	204254_s_at
VEGFA	**-**	**37**	**-**	**20**	**47**	**-**	**-**	**-**	**54**	**74**	**52**	**76**	**63**	**210512_s_at**
	97	**13**	**43**	**10**	**10**	**93**	**-**	**-**	**51**	**37**	**27**	**29**	**19**	**210513_s_at**
	-	**79**	**-**	**100**	**21**	**-**	**-**	**-**	**39**	**39**	**41**	**44**	**50**	**211527_x_at**
	-	**34**	**-**	**19**	**19**	**-**	**-**	**-**	**38**	**32**	**33**	**37**	**31**	**212171_x_at**
VGLL4	*10*	*23*	*46*	*69*	*23*	*9*	*2*	*21*	*75*	*8*	*15*	*4*	*11*	*212399_s_at*
	-	-	-	*73*	*78*	-	-	-	-	-	-	-	*87*	*214004_s_at*
VIM	85	-	-	-	-	92	-	-	-	-	-	-	-	201426_s_at
VPS35	-	-	-	-	-	-	24	-	-	-	-	-	-	217727_x_at
WSB1	**-**	**61**	**-**	**43**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**89**	**201295_s_at**
WWC3	-	-	-	95	-	-	-	-	-	-	-	-	-	219520_s_at
ZBTB1	32	-	16	-	-	39	-	-	-	-	-	-	-	213376_at
ZBTB17	-	-	-	-	-	-	100	32	-	-	-	-	-	203601_s_at
ZBTB43	-	-	-	-	-	-	81	26	-	-	-	-	-	204181_s_at
ZFR	-	-	-	-	-	-	13	82	-	-	-	-	-	213286_at
ZMPSTE24	74	-	64	-	-	73	-	-	-	-	-	-	-	202939_at
ZNF274	-	-	-	-	-	-	87	39	-	-	-	-	-	204937_s_at
ZNF395	-	*17*	-	*58*	*74*	-	-	-	*27*	*63*	*45*	*69*	*58*	*221123_x_at*
	25	*3*	*73*	*26*	*25*	*23*	-	-	*1*	*1*	*1*	*5*	*5*	*218149_s_at*
ZNHIT3	69	-	56	-	-	71	-	-	-	-	-	-	-	212544_at
ZWINT	-	-	-	92	-	-	-	-	-	-	-	-	-	204026_s_at

Annexe I - fin: Analyse du jeu de données E-MEXP-445 : sélection des 100 *probesets* les plus significatifs pour chaque méthode. Pour chaque *probeset* sélectionné, la table fournie présente le rang du *probeset* dans chacune des *top-lists* définies par les méthodes individuelles. En gras : gènes indiqués listés dans l'étude originale. En gris sur-ligné : gènes validés dans d'autres études. En italique souligné : gènes candidats pour une étude ultérieure.

Annexe II

Bibliographie relative aux gènes validés dans un contexte hypoxique

ADAM8	34
ADFP (ADPH)	28
c3orf28 (HGTD-P)	16
CCL18	27
CD163	4, 5, 19
CHI3L1	13, 25, 29
CXCL5	5
CXCR7	30, 8
DAPK1	35
ELK1	20
ENPP2 (ATX)	5, 3
FCGR1A (CD64), FCGR1B (CD64), FCGR2B (CD32), FCGR2C (CD32), FCAR (CD89)	4
GPR109B (HM74)	15
LONP1 (LON,PDK1)	31, 9, 12
MAPK7 (BMK1,ERK5)	23, 24
MMP19	4, 5
NT5E (CD73)	18, 33, 32, 2
PLOD2	6, 11
PRDX3	36, 7, 22
TFPI2 (PP5)	37, 21, 10
TMEM45A	26
TXNIP	1, 17, 14

1. Baker A.F., Koh M.Y., Williams R.R., James B., Wang H., Tate W.R. *et al.* Identification of thioredoxin-interacting protein 1 as a hypoxia-inducible factor 1alpha-induced gene in pancreatic cancer, *Pancreas*, 2008, 36(2) : 178-186.

2. Berchtold S., Ogilvie A.L., Bogdan C., Mühl-Zürbes P., Ogilvie A., Schuler G. *et al.* Human monocyte derived dendritic cells express functional P2X and P2Y receptors as well as ecto-nucleotidases, *FEBS Lett.*, 1999, 458(3) : 424-428.

3. Black E.J., Clair T., Delrow J., Neiman P., Gillespie D.A. Microarray analysis identifies Autotaxin, a tumour cell motility and angiogenic factor with lysophospholipase D activity, as a specific target of cell transformation by v-Jun. *Oncogene*, 2004, 23(13) : 2357-2366.

4. Bosco M.C., Puppo M., Santangelo C., Anfosso L., Pfeffer U., Fardin P. *et al.* Hypoxia modifies

the transcriptome of primary human monocytes: modulation of novel immune-related genes and identification of CC-chemokine ligand 20 as a new hypoxia-inducible gene. *J. Immunol.*, 2006, 177(3) : 1941-55.

5. Bosco M.C., Puppo M., Blengio F., Fraone T., Cappello P., Giovarelli M. *et al.* Monocytes and dendritic cells in a hypoxic environment: Spotlights on chemotaxis and migration. *Immunobiology*, 2008, 13(9-10) : 733-49.

6. Brinckmann J., Kim S., Wu J., Reinhardt D.P., Batmunkh C., Metzen E. *et al.*, Interleukin 4 and prolonged hypoxia induce a higher gene expression of lysyl hydroxylase 2 and an altered cross-link pattern: important pathogenetic steps in early and late stage of systemic scleroderma? *Matrix Biol.*, 2005, 24(7) : 459-468.

7. Chang T.S., Cho C.S., Park S., Yu S., Kang S.W., Rhee S.G., Peroxiredoxin III, a mitochondrion-specific peroxidase, regulates apoptotic signaling by mitochondria, *J. Biol. Chem.*, 2004, 279(40) : 41975-41984.

8. Costello C.M., Howell K., Cahill E., McBryan J., Konigshoff M., Eickelberg O. *et al.*, Lung-selective gene responses to alveolar hypoxia: potential role for the bone morphogenetic antagonist gremlin in pulmonary hypertension., *Am. J. Physiol. Lung C.*, 2008, 295(2) : L272-284.

9. Fukuda R., Zhang H., Kim J.W., Shimoda L., Dang C.V., Semenza G.L., HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells., *Cell.* 2007, 129(1) : 111-122.

10. Golden T., Aragon I.V., Rutland B., Tucker J.A., Shevde L.A., Samant R.S. *et al.*, Honkanen R.E. Elevated levels of Ser/Thr protein phosphatase 5 (PP5) in human breast cancer, *Biochim. Biophys. Acta.*, 2008, 782(4) : 259-270.

11. Hofbauer K.H., Gess B., Lohaus C., Meyer H.E., Katschinski D., Kurtz A., Oxygen tension regulates the expression of a group of procollagen hydroxylases, *Eur. J. Biochem.*, 2003, 270(22) : 4515-4522.

12. Hori O., Ichinoda F., Tamatani T., Yamaguchi A., Sato N., Ozawa K. *et al.*, Transmission of cell stress from endoplasmic reticulum to mitochondria: enhanced expression of Lon protease. *J. Cell. Biol.*, 2002, 57(7) : 1151-1160.

13. Junker N., Johansen J.S., Hansen L.T., Lund E.L., Kristjansen P.E., Regulation of YKL-40 expression during genotoxic or microenvironmental stress in human glioblastoma cells, *Cancer Sci.*, 2005, 96(3) : 183-90.

14. Karar J., Dolt K.S., Mishra M.K., Arif E., Javed S., Pasha M.A., Expression and functional activity of pro-oxidants and antioxidants in murine heart exposed to acute hypobaric hypoxia, *FEBS Lett.*, 2007, 581(24) : 4577-4582.

15. Knowles H.J., te Poele R.H., Workman P., Harris A.L., Niacin induces PPARgamma expression and transcriptional activation in macrophages via HM74 and HM74a-mediated induction of prostaglandin synthesis pathways., *Biochem. Pharmacol.*, 2006, 71(5) : 646-56.

16. Lee M.J., Kim J.Y., Suk K., Park J.H., Identification of the Hypoxia-Inducible Factor 1α-Responsive HGTD-P Gene as a Mediator in the Mitochondrial Apoptotic Pathway, *Mol. Cell Biol.*, 2004, 24(9) : 3918-27.

17. Le Jan S., Le Meur N., Cazes A., Philippe J., Le Cunff M., Léger J. *et al.*, Characterization of the expression of the hypoxia-induced genes neuritin, TXNIP and IGFBP3 in cancer, *FEBS Lett.*,

2006, 580(14) : 3395-3400.

18. Li X., Zhou T., Zhi X., Zhao F., Yin L., Zhou P., Effect of hypoxia/reoxygenation on CD73 (ecto-5'-nucleotidase) in mouse microvessel endothelial cell lines. *Microvasc. Res.*, 2006, 72(1-2) : 48-53.

19. Moniuszko M., Kowal K., Rusak M., Pietruczuk M., Dabrowska M., Bodzenta-Lukaszyk A. Monocyte CD163 and CD36 expression in human whole blood and isolated mononuclear cell samples: influence of different anticoagulants. *Clin. Vaccine Immunol.*, 2006, 13(6) : 704-7.

20. Müller J.M., Krauss B., Kaltschmidt C., Baeuerle P.A., Rupec R.A., Hypoxia induces c-fos transcription via a mitogen-activated protein kinase-dependent pathway., *J. Biol. Chem.*, 1997, 272(37) : 23435-23439.

21. Ni L., Swingle M.S., Bourgeois A.C., Honkanen R.E., High yield expression of serine/threonine protein phosphatase type 5, and a fluorescent assay suitable for use in the detection of catalytic inhibitors, *Assay Drug Dev. Techn.*, 2007, 5(5) : 645-653.

22. Nonn L., Berggren M., Powis G., Increased expression of mitochondrial peroxiredoxin-3 (thioredoxin peroxidase-2) protects cancer cells against hypoxia and drug-induced hydrogen peroxide-dependent apoptosis, *Mol. Cancer Res.*, 2003, 1(9) : 682-689.

23. Pi X., Yan C., Berk B.C., Big mitogen-activated protein kinase (BMK1)/ERK5 protects endothelial cells from apoptosis. *Circ. Res.*, 2004, 94(3) : 362-369.

24. Pi X., Garin G., Xie L., Zheng Q., Wei H., Abe J. *et al.*, BMK1/ERK5 is a novel regulator of angiogenesis by destabilizing hypoxia inducible factor 1alpha., *Circ. Res.*, 2005, 96(11) : 1145-1151.

25. Recklies A.D., Ling H., White C., Bernier S.M. Inflammatory cytokines induce production of CHI3L1 by articular chondrocytes. *J. Biol. Chem.*, 2005, 280(50) : 41213-21.

26. Rendon E., Hale S.J., Ryan D., Baban D., Forde S.P., Roubelakis M. *et al.*, Transcriptional profiling of human cord blood CD133+ and cultured bone marrow mesenchymal stem cells in response to hypoxia, *Stem Cells.*, 2007, 25(4) : 1003-1012.

27. Ricciardi A., Elia A.R., Cappello P., Puppo M., Vanni C., Fardin P. *et al.*, Transcriptome of hypoxic immature dendritic cells: modulation of chemokine/receptor expression, *Mol. Cancer Res.*, 2008, 6(2) : 175-85.

28. Saarikoski S.T., Rivera S.P., Hankinson O., Mitogen-inducible gene 6 (MIG-6), adipophilin and tuftelin are inducible by hypoxia, *FEBS Lett.*, 2002, 530(1-3) : 186-90.

29. Saidi A., Javerzat S., Bellahcène A., De Vos J., Bello L., Castronovo V. *et al.*, Experimental anti-angiogenesis causes upregulation of genes associated with poor survival in glioblastoma., *Int. J. Cancer.*, 2008, 122(10) : 2187-98.

30. Schutyser E., Su Y., Yu Y., Gouwy M., Zaja-Milatovic S., Van Damme J. *et al.*, Hypoxia enhances CXCR4 expression in human microvascular endothelial cells and human melanoma cells., *Eur. Cytokine Netw.*, 2007, 59-70.

31. Semenza G.L., Oxygen-dependent regulation of mitochondrial respiration by hypoxia-inducible factor 1., *Biochem. J.*, 2007, 405(1) : 1-9.

32. Synnestvedt K., Furuta G.T., Comerford K.M., Louis N., Karhausen J., Eltzschig H.K. *et al.*, Colgan S.P., Ecto-5'-nucleotidase (CD73) regulation by hypoxia-inducible factor-1 mediates permeability changes in intestinal epithelia. *J. Clin. Invest.* 2002, 110(7) : 993-1002.

33. Thompson L.F., Eltzschig H.K., Ibla J.C., Van De Wiele C.J., Resta R., Morote-Garcia J.C., Colgan S.P. Crucial role for ecto-5'-nucleotidase (CD73) in vascular leakage during hypoxia., *J.*

Exp. Med., 2004, 200(11) : 1395-405.

34. Valkovskaya N.V. Hypoxia-dependent expression of ADAM8 in human pancreatic cancer cell lines, *Exp. Oncol.*, 2008, 30(2) :129-32.

35. Velentza A.V., Wainwright M.S., Zasadzki M., Mirzoeva S., Schumacher A.M., Haiech J. *et al.*, An aminopyridazine-based inhibitor of a pro-apoptotic protein kinase attenuates hypoxia-ischemia induced acute brain injury, *Bioorg. Med. Chem. Lett.*, 2003, 13(20) : 3465-3470.

36. Zhang H., Go Y.M., Jones D.P. Mitochondrial thioredoxin-2/peroxiredoxin-3 system functions in parallel with mitochondrial GSH system in protection against oxidative stress, *Arch. Biochem. Biophys.*, 2007, 465(1) : 119-126.

37. Zhou G., Golden T., Aragon I.V., Honkanen R.E. Ser/Thr protein phosphatase 5 inactivates hypoxia-induced activation of an apoptosis signal-regulating kinase 1/MKK-4/JNK signaling cascade, *J. Biol. Chem.*, 2004, 279(45) : 46595-605.

Source	Méthode	Groupes
C2.biocarta	a2.fixed	actinypathway,chemicalpathway,crebpathway,erythpathway,feederpathway,glycolysispathway,malatexpathway,rhopathway,srcrptppathway,tall1pathway,vitcbpathway
	faeri.fixed.perms	raspathway
	GSA.mean.*	feederpathway,ghpathway,glycolysispathway,hifpathway,igf1pathway,plcpathway,sarspathway,vitcbpathway
	GSA.absmean.*	feederpathway,glycolysispathway
	GSA.maxmean.*	actinypathway,chemicalpathway,feederpathway,glycolysispathway,nkcellspathway,p53hypoxiapathway,plcpathway,sarspathway,vitcbpathway
	globaltest.gamma	blymphocytepathway,feederpathway,glycolysispathway,hifpathway,no1pathway,ptdinspathway
C2.kegg	a2.fixed	hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00061_fatty_acid_biosynthesis,hsa00071_fatty_acid_metabolism,hsa00072_synthesis_and_degradation_of_ketone_bodies,hsa00100_biosynthesis_of_steroids,hsa00190_oxidative_phosphorylation,hsa00232_caffeine_metabolism,hsa00330_arginine_and_proline_metabolism,hsa00380_tryptophan_metabolism,hsa00480_glutathione_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa00620_pyruvate_metabolism,hsa00630_glyoxylate_and_dicarboxylate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa00740_riboflavin_metabolism,hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa03320_ppar_signaling_pathway,hsa04150_mtor_signaling_pathway,hsa04370_vegf_signaling_pathway,hsa04530_tight_junction,hsa04612_antigen_processing_and_presentation,hsa05030_amyotrophic_lateral_sclerosis,hsa05040_huntingtons_disease,hsa05110_cholera_infection,hsa05120_epithelial_cell_signaling_in_helicobacter_pylori_infection,hsa05211_renal_cell_carcinoma
	faeri.fixed.null	hsa00140_c21_steroid_hormone_metabolism,hsa04940_type_i_diabetes_mellitus,hsa04950_maturity_onset_diabetes_of_the_young,hsa05210_colorectal_cancer,hsa05211_renal_cell_carcinoma,hsa05212_pancreatic_cancer,hsa05213_endometrial_cancer
	faeri.fixed.perms	hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00100_biosynthesis_of_steroids,hsa00511_n_glycan_degradation,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa04130_snare_interactions_in_vesicular_transport,hsa04210_apoptosis,hsa04620_toll_like_receptor_signaling_pathway,hsa04662_b_cell_receptor_signaling_pathway,hsa05040_huntingtons_disease,hsa05216_thyroid_cancer,hsa05221_acute_myeloid_leukemia
	GSA.mean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00052_galactose_metabolism,hsa00053_ascorbate_and_aldarate_metabolism,hsa00071_fatty_acid_metabolism,hsa00340_histidine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00710_carbon_fixation,hsa04660_t_cell_receptor_signaling_pathway,hsa05010_alzheimers_disease,hsa05050_dentatorubropallidoluysian_atrophy
	GSA.absmean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00031_inositol_metabolism,hsa00521_streptomycin_biosynthesis,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy
	GSA.maxmean.*	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00100_biosynthesis_of_steroids,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00640_propanoate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa01510_neurodegenerative_diseases,hsa04660_t_cell_receptor_signaling_pathway,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05010_alzheimers_disease,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy,hsa05211_renal_cell_carcinoma
	globaltest.gamma	hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00052_galactose_metabolism,hsa00272_cysteine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05211_renal_cell_carcinoma

Annexe III - page 1

Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Pour chacune des méthodes utilisées, la seuil de sélection a été placé à 0,01 sur les *p-values*.

Source	Méthode	Groupes
C2.genmapp	a2.fixed	atp_synthesis,bile_acid_biosynthesis,biosynthesis_of_steroids,carbon_fixation,cholesterol_biosynthesis,citrate_cycle_tca_cycle,flagellar_assembly,folate_biosynthesis,fructose_and_mannose_metabolism,galactose_metabolism,gluconeogenesis,glutathione_metabolism,glycolysis,glycolysis_and_gluconeogenesis,glycosphingolipid_metabolism,glyoxylate_and_dicarboxylate_metabolism,inositol_metabolism,oxidative_phosphorylation,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,photosynthesis,pyruvate_metabolism,reductive_carboxylate_cycle_co2_fixation,riboflavin_metabolism,ribosomal_proteins,starch_and_sucrose_metabolism,streptomycin_biosynthesis,synthesis_and_degradation_of_ketone_bodies,terpenoid_biosynthesis,tryptophan_metabolism,type_iii_secretion_system
	faeri.fixed.null	heparan_sulfate_biosynthesis,limonene_and_pinene_degradation,type_iii_secretion_system,tyrosine_metabolism
	faeri.fixed.perms	biosynthesis_of_steroids,cholesterol_biosynthesis,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,pentose_phosphate_pathway,pyruvate_metabolism,ribosomal_proteins,terpenoid_biosynthesis
	GSA.mean.*	bile_acid_biosynthesis,carbon_fixation,fructose_and_mannose_metabolism,galactose_metabolism,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,inositol_metabolism,o_glycan_biosynthesis,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,riboflavin_metabolism,smooth_muscle_contraction,starch_and_sucrose_metabolism,streptomycin_biosynthesis
	GSA.absmean.*	gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,inositol_metabolism,phenylalanine_tyrosine_and_tryptophan_biosynthesis,streptomycin_biosynthesis
	GSA.maxmean.*	aminosugars_metabolism,carbon_fixation,cholesterol_biosynthesis,fructose_and_mannose_metabolism,galactose_metabolism,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,inositol_metabolism,mitochondrial_fatty_acid_betaoxidation,o_glycan_biosynthesis,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,photosynthesis,propanoate_metabolism,reductive_carboxylate_cycle_co2_fixation,riboflavin_metabolism,starch_and_sucrose_metabolism,streptomycin_biosynthesis,terpenoid_biosynthesis,ubiquitin_mediated_proteolysis
	globaltest.gamma	cysteine_metabolism,galactose_metabolism,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,gpcrdb_class_a_rhodopsin_like2,krebs_tca_cycle,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,starch_and_sucrose_metabolism,streptomycin_biosynthesis
C2.canonical	a2.fixed	actinypathway,atp_synthesis,bile_acid_biosynthesis,biosynthesis_of_steroids,carbon_fixation,chemicalpathway,cholesterol_biosynthesis,citrate_cycle_tca_cycle,crebpathway,erythpathway,feederpathway,flagellar_assembly,folate_biosynthesis,fructose_and_mannose_metabolism,galactose_metabolism,gluconeogenesis,glutathione_metabolism,glycolysis,glycolysis_and_gluconeogenesis,glycolysispathway,glycosphingolipid_metabolism,glyoxylate_and_dicarboxylate_metabolism,hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00061_fatty_acid_biosynthesis,hsa00071_fatty_acid_metabolism,hsa00072_synthesis_and_degradation_of_ketone_bodies,hsa00100_biosynthesis_of_steroids,hsa00190_oxidative_phosphorylation,hsa00232_caffeine_metabolism,hsa00330_arginine_and_proline_metabolism,hsa00380_tryptophan_metabolism,hsa00480_glutathione_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa00620_pyruvate_metabolism,hsa00630_glyoxylate_and_dicarboxylate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa00740_riboflavin_metabolism,hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa03320_ppar_signaling_pathway,hsa04150_mtor_signaling_pathway,hsa04370_vegf_signaling_pathway,hsa04530_tight_junction,hsa04612_antigen_processing_and_presentation,hsa05030_amyotrophic_lateral_sclerosis,hsa05040_huntingtons_disease,hsa05110_cholera_infection,hsa05120_epithelial_cell_signaling_in_helicobacter_pylori_infection,hsa05211_renal_cell_carcinoma,inositol_metabolism,malatexpathway,oxidative_phosphorylation,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,photosynthesis,pyruvate_metabolism,reductive_carboxylate_cycle_co2_fixation,rhopathway,riboflavin_metabolism,ribosomal_proteins,srcrptppathway,starch_and_sucrose_metabolism,streptomycin_biosynthesis,synthesis_and_degradation_of_ketone_bodies,tall1pathway,terpenoid_biosynthesis,tryptophan_metabolism,type_iii_secretion_system,vitcbpathway
	faeri.fixed.null	heparan_sulfate_biosynthesis,hsa00140_c21_steroid_hormone_metabolism,hsa04940_type_i_diabetes_mellitus,hsa04950_maturity_onset_diabetes_of_the_young,hsa05210_colorectal_cancer,hsa05211_renal_cell_carcinoma,hsa05212_pancreatic_cancer,hsa05213_endometrial_cancer,limonene_and_pinene_degradation,type_iii_secretion_system,tyrosine_metabolism
	faeri.fixed.perms	biosynthesis_of_steroids,cholesterol_biosynthesis,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,hsa00010_glycolysis_and_gluconeogenesis,hsa00020_citrate_cycle,hsa00030_pentose_phosphate_pathway,hsa00100_biosynthesis_of_steroids,hsa00511_n_glycan_degradation,hsa01032_glycan_structures_degradation,hsa03010_ribosome,hsa04130_snare_interactions_in_vesicular_transport,hsa04210_apoptosis,hsa04620_toll_like_receptor_signaling_pathway,hsa04662_b_cell_receptor_signaling_pathway,hsa05040_huntingtons_disease,hsa05216_thyroid_cancer,hsa05221_acute_myeloid_leukemia,pentose_phosphate_pathway,pyruvate_metabolism,raspathway,ribosomal_proteins,sa_b_cell_receptor_complexes,terpenoid_biosynthesis

Annexe III - page 2

Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Pour chacune des méthodes utilisées, la seuil de sélection a été placé à 0,01 sur les *p-values*.

Source	Méthode	Groupes
C2.canonical	GSA.mean.*	bile_acid_biosynthesis,carbon_fixation,feederpathway,fructose_and_mannose_metabolism,galactose_metabolism,ghpathway,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,glycolysispathway,hifpathway,hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00052_galactose_metabolism,hsa00053_ascorbate_and_aldarate_metabolism,hsa00071_fatty_acid_metabolism,hsa00340_histidine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00710_carbon_fixation,hsa04660_t_cell_receptor_signaling_pathway,hsa05010_alzheimers_disease,hsa05050_dentatorubropallidoluysian_atrophy,igf1pathway,inositol_metabolism,o_glycan_biosynthesis,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,plcpathway,riboflavin_metabolism,sa_pten_pathway,sarspathway,sig_il4receptor_in_b_lyphocytes,sig_pip3_signaling_in_cardiac_myoctes,smooth_muscle_contraction,st_g_alpha_i_pathway,st_ga13_pathway,starch_and_sucrose_metabolism,streptomycin_biosynthesis,vitcbpathway
	GSA.absmean.*	feederpathway,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,glycolysispathway,hsa00010_glycolysis_and_gluconeogenesis,hsa00031_inositol_metabolism,hsa00521_streptomycin_biosynthesis,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy,inositol_metabolism,phenylalanine_tyrosine_and_tryptophan_biosynthesis,streptomycin_biosynthesis
	GSA.maxmean.*	actinypathway,aminosugars_metabolism,carbon_fixation,chemicalpathway,cholesterol_biosynthesis,feederpathway,fructose_and_mannose_metabolism,galactose_metabolism,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,glycolysispathway,hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00031_inositol_metabolism,hsa00051_fructose_and_mannose_metabolism,hsa00052_galactose_metabolism,hsa00100_biosynthesis_of_steroids,hsa00500_starch_and_sucrose_metabolism,hsa00512_o_glycan_biosynthesis,hsa00521_streptomycin_biosynthesis,hsa00640_propanoate_metabolism,hsa00710_carbon_fixation,hsa00720_reductive_carboxylate_cycle,hsa01510_neurodegenerative_diseases,hsa04660_t_cell_receptor_signaling_pathway,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05010_alzheimers_disease,hsa05040_huntingtons_disease,hsa05050_dentatorubropallidoluysian_atrophy,hsa05211_renal_cell_carcinoma,inositol_metabolism,mitochondrial_fatty_acid_betaoxidation,nkcellspathway,o_glycan_biosynthesis,p53hypoxiapathway,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,photosynthesis,plcpathway,propanoate_metabolism,reductive_carboxylate_cycle_co2_fixation,riboflavin_metabolism,sarspathway,sig_il4receptor_in_b_lyphocytes,sig_insulin_receptor_pathway_in_cardiac_myocytes,sig_pip3_signaling_in_b_lymphocytes,st_ga13_pathway,st_gaq_pathway,st_pac1_receptor_pathway,starch_and_sucrose_metabolism,streptomycin_biosynthesis,terpenoid_biosynthesis,ubiquitin_mediated_proteolysis,vitcbpathway
	globaltest.gamma	blymphocytepathway,cysteine_metabolism,feederpathway,galactose_metabolism,gluconeogenesis,glycolysis,glycolysis_and_gluconeogenesis,glycolysispathway,gpcrdb_class_a_rhodopsin_like2,hifpathway,hsa00010_glycolysis_and_gluconeogenesis,hsa00030_pentose_phosphate_pathway,hsa00052_galactose_metabolism,hsa00272_cysteine_metabolism,hsa00500_starch_and_sucrose_metabolism,hsa00521_streptomycin_biosynthesis,hsa04664_fc_epsilon_ri_signaling_pathway,hsa05211_renal_cell_carcinoma,krebs_tca_cycle,no1pathway,pentose_phosphate_pathway,phenylalanine_tyrosine_and_tryptophan_biosynthesis,ptdinspathway,sa_pten_pathway,sig_bcr_signaling_pathway,sig_chemotaxis,sig_il4receptor_in_b_lyphocytes,sig_insulin_receptor_pathway_in_cardiac_myocytes,sig_pip3_signaling_in_b_lymphocytes,sig_pip3_signaling_in_cardiac_myoctes,st_ga13_pathway,st_gaq_pathway,st_phosphoinositide_3_kinase_pathway,starch_and_sucrose_metabolism,streptomycin_biosynthesis

Annexe III - fin

Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Pour chacune des méthodes utilisées, la seuil de sélection a été placé à 0,01 sur les *p-values*.

a2.fixed	hsa00010_glycolysis_and_gluconeogenesis, hsa00020_citrate_cycle, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00100_biosynthesis_of_steroids, hsa00190_oxidative_phosphorylation, hsa00330_arginine_and_proline_metabolism, hsa00480_glutathione_metabolism, hsa00500_starch_and_sucrose_metabolism, hsa00521_streptomycin_biosynthesis, hsa00620_pyruvate_metabolism, hsa00720_reductive_carboxylate_cycle, hsa00740_riboflavin_metabolism, hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450, hsa03010_ribosome, hsa04612_antigen_processing_and_presentation, hsa05120_epithelial_cell_signaling_in_helicobacter_pylori_infection, hsa05211_renal_cell_carcinoma
faeri.fixed.null	hsa00031_inositol_metabolism, hsa00140_c21_steroid_hormone_metabolism, hsa00300_lysine_biosynthesis, hsa00532_chondroitin_sulfate_biosynthesis, hsa00860_porphyrin_and_chlorophyll_metabolism, hsa00900_terpenoid_biosynthesis, hsa00970_aminoacyl_trna_biosynthesis, hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450, hsa04110_cell_cycle, hsa04115_p53_signaling_pathway, hsa04210_apoptosis, hsa04310_wnt_signaling_pathway, hsa04614_renin_angiotensin_system, hsa04620_toll_like_receptor_signaling_pathway, hsa04940_type_i_diabetes_mellitus, hsa04950_maturity_onset_diabetes_of_the_young, hsa05210_colorectal_cancer, hsa05211_renal_cell_carcinoma, hsa05212_pancreatic_cancer, hsa05213_endometrial_cancer
faeri.fixed.perms	hsa00010_glycolysis_and_gluconeogenesis, hsa00020_citrate_cycle, hsa00030_pentose_phosphate_pathway, hsa00051_fructose_and_mannose_metabolism, hsa00100_biosynthesis_of_steroids, hsa00511_n_glycan_degradation, hsa00620_pyruvate_metabolism, hsa01032_glycan_structures_degradation, hsa01510_neurodegenerative_diseases, hsa03010_ribosome, hsa03320_ppar_signaling_pathway, hsa04130_snare_interactions_in_vesicular_transport, hsa04210_apoptosis, hsa04620_toll_like_receptor_signaling_pathway, hsa04650_natural_killer_cell_mediated_cytotoxicity, hsa04662_b_cell_receptor_signaling_pathway, hsa05040_huntingtons_disease, hsa05212_pancreatic_cancer, hsa05216_thyroid_cancer, hsa05221_acute_myeloid_leukemia
GSA.mean.*	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00053_ascorbate_and_aldarate_metabolism, hsa00071_fatty_acid_metabolism, hsa00120_bile_acid_biosynthesis, hsa00340_histidine_metabolism, hsa00380_tryptophan_metabolism, hsa00500_starch_and_sucrose_metabolism, hsa00512_o_glycan_biosynthesis, hsa00521_streptomycin_biosynthesis, hsa00626_naphthalene_and_anthracene_degradation, hsa00710_carbon_fixation, hsa00950_alkaloid_biosynthesis_i, hsa01040_polyunsaturated_fatty_acid_biosynthesis, hsa01510_neurodegenerative_diseases, hsa04660_t_cell_receptor_signaling_pathway, hsa04664_fc_epsilon_ri_signaling_pathway, hsa05010_alzheimers_disease, hsa05050_dentatorubropallidoluysian_atrophy
GSA.absmean.*	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00140_c21_steroid_hormone_metabolism, hsa00430_taurine_and_hypotaurine_metabolism, hsa00521_streptomycin_biosynthesis, hsa00591_linoleic_acid_metabolism, hsa00710_carbon_fixation, hsa01430_cell_communication, hsa01510_neurodegenerative_diseases, hsa04020_calcium_signaling_pathway, hsa04080_neuroactive_ligand_receptor_interaction, hsa04140_regulation_of_autophagy, hsa04340_hedgehog_signaling_pathway, hsa04530_tight_junction, hsa04614_renin_angiotensin_system, hsa04720_long_term_potentiation, hsa04740_olfactory_transduction, hsa04742_taste_transduction, hsa04950_maturity_onset_diabetes_of_the_young, hsa05010_alzheimers_disease, hsa05040_huntingtons_disease, hsa05050_dentatorubropallidoluysian_atrophy
GSA.maxmean.*	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00100_biosynthesis_of_steroids, hsa00500_starch_and_sucrose_metabolism, hsa00512_o_glycan_biosynthesis, hsa00521_streptomycin_biosynthesis, hsa00564_glycerophospholipid_metabolism, hsa00640_propanoate_metabolism, hsa00710_carbon_fixation, hsa00720_reductive_carboxylate_cycle, hsa01510_neurodegenerative_diseases, hsa04660_t_cell_receptor_signaling_pathway, hsa04664_fc_epsilon_ri_signaling_pathway, hsa05010_alzheimers_disease, hsa05040_huntingtons_disease, hsa05050_dentatorubropallidoluysian_atrophy, hsa05211_renal_cell_carcinoma

Annexe IV - page 1

Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Pour chaque méthode, les 20 groupes les plus significatifs sont listés.

globaltest.asymptotic	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00061_fatty_acid_biosynthesis, hsa00232_caffeine_metabolism, hsa00272_cysteine_metabolism, hsa00330_arginine_and_proline_metabolism, hsa00350_tyrosine_metabolism, hsa00360_phenylalanine_metabolism, hsa00401_novobiocin_biosynthesis, hsa00500_starch_and_sucrose_metabolism, hsa00521_streptomycin_biosynthesis, hsa00565_ether_lipid_metabolism, hsa00950_alkaloid_biosynthesis_i, hsa04150_mtor_signaling_pathway, hsa04664_fc_epsilon_ri_signaling_pathway, hsa04710_circadian_rhythm, hsa05211_renal_cell_carcinoma
globaltest.gamma	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00272_cysteine_metabolism, hsa00330_arginine_and_proline_metabolism, hsa00350_tyrosine_metabolism, hsa00360_phenylalanine_metabolism, hsa00500_starch_and_sucrose_metabolism, hsa00521_streptomycin_biosynthesis, hsa00562_inositol_phosphate_metabolism, hsa00565_ether_lipid_metabolism, hsa04150_mtor_signaling_pathway, hsa04370_vegf_signaling_pathway, hsa04662_b_cell_receptor_signaling_pathway, hsa04664_fc_epsilon_ri_signaling_pathway, hsa04710_circadian_rhythm, hsa05110_cholera_infection, hsa05211_renal_cell_carcinoma
gsea.pval	hsa00140_c21_steroid_hormone_metabolism, hsa00220_urea_cycle_and_metabolism_of_amino_groups, hsa00232_caffeine_metabolism, hsa00601_glycosphingolipid_biosynthesis_lactoseries, hsa00620_pyruvate_metabolism, hsa00750_vitamin_b6_metabolism, hsa00770_pantothenate_and_coa_biosynthesis, hsa01032_glycan_structures_degradation, hsa01040_polyunsaturated_fatty_acid_biosynthesis, hsa03320_ppar_signaling_pathway, hsa04020_calcium_signaling_pathway, hsa04060_cytokine_cytokine_receptor_interaction, hsa04080_neuroactive_ligand_receptor_interaction, hsa04370_vegf_signaling_pathway, hsa04540_gap_junction, hsa04710_circadian_rhythm, hsa04720_long_term_potentiation, hsa04740_olfactory_transduction, hsa04920_adipocytokine_signaling_pathway, hsa05030_amyotrophic_lateral_sclerosis
gsea.fdr	hsa00010_glycolysis_and_gluconeogenesis, hsa00052_galactose_metabolism, hsa00061_fatty_acid_biosynthesis, hsa00232_caffeine_metabolism, hsa00340_histidine_metabolism, hsa00361_gamma_hexachlorocyclohexane_degradation, hsa00500_starch_and_sucrose_metabolism, hsa00521_streptomycin_biosynthesis, hsa00602_glycosphingolipid_biosynthesis_neo_lactoseries, hsa00620_pyruvate_metabolism, hsa00630_glyoxylate_and_dicarboxylate_metabolism, hsa00641_3_chloroacrylic_acid_degradation, hsa00940_phenylpropanoid_biosynthesis, hsa00980_metabolism_of_xenobiotics_by_cytochrome_p450, hsa01430_cell_communication, hsa02010_abc_transporters_general, hsa04530_tight_junction, hsa04540_gap_junction, hsa04612_antigen_processing_and_presentation, hsa04742_taste_transduction
samgs.pval	hsa00010_glycolysis_and_gluconeogenesis, hsa00030_pentose_phosphate_pathway, hsa00031_inositol_metabolism, hsa00051_fructose_and_mannose_metabolism, hsa00052_galactose_metabolism, hsa00061_fatty_acid_biosynthesis, hsa00100_biosynthesis_of_steroids, hsa00330_arginine_and_proline_metabolism, hsa00401_novobiocin_biosynthesis, hsa00500_starch_and_sucrose_metabolism, hsa00512_o_glycan_biosynthesis, hsa00521_streptomycin_biosynthesis, hsa00620_pyruvate_metabolism, hsa00710_carbon_fixation, hsa00720_reductive_carboxylate_cycle, hsa00900_terpenoid_biosynthesis, hsa01510_neurodegenerative_diseases, hsa04660_t_cell_receptor_signaling_pathway, hsa04664_fc_epsilon_ri_signaling_pathway, hsa04710_circadian_rhythm

Annexe IV - fin

Liste des groupes détectés par les différentes méthodes d'analyse, sur le jeux de données E-MEXP-445. Les catégories représentées concernent uniquement la définition des voies métaboliques. Pour chaque méthode, les 20 groupes les plus significatifs sont listés. GlobalTest.permutations et SAMGS.qval ne sont pas représentés en raison du grand nombre de groupes qui obtiennent le même score.